浙江工商大学 · 通识选修课教材

人工智能辅助的数字媒体创作

原理图解 · 案例教学 · 参数实操

从一句提示词到一部作品：一条流水线走完文本、声音、图像、视频、三维、动画、姿态、场景与智能体。

贯穿案例：三人学生小组用八章所学完成古诗词短片《春江花月夜》。

导论 + 8 章 + 6 个附录 | 124 幅手绘 SVG 图解 + 20 幅 AI 生成插图 | 配套 32 学时课程（16 讲授 + 16 实践）

2026 年 · 浙江工商大学

目　录

导论 · 当生成变得廉价，创作还剩下什么 4

0.1 为什么是现在：三个绕不开的事实 4
0.2 一张地图：八种媒介，一条流水线 5
0.3 机器是怎么学会"创作"的：三次浪潮与一个拐点 6
0.4 能力边界：它擅长什么、又做不好什么 7
0.5 这本书怎么用：八段体例与四条轨道 8
0.6 期末你会交出什么：评分即路线图 9
0.7 工具、账号与三条合规底线 10
0.8 人在回路：这门课真正想教的东西 11

第 1 章 · 人工智能与多模态基础 · 大语言模型与声音生成 13

1.1 多模态：机器眼中的世界，全都是数字 13
1.2 大语言模型：一台"预测下一个词"的机器 16
1.3 推理模型：从"脱口而出"到"想好再说" 19
1.4 提示工程：把"想要什么"说清楚的手艺 20
1.5 幻觉：模型什么时候会"一本正经地胡说" 23
1.6 语音识别（STT）：把声音变成文字 25
1.7 语音合成（TTS）与语音克隆：把文字变成人声 27
1.8 音乐与音效生成：第四条产线 30
1.9 工具地图与合规框架 31
1.10 关键概念表 35
1.11 动手实践：从环境准备到第一段配音 37
1.12 课程思政：文化传播与声音权益 38

第 2 章 · 图像生成与编辑 45

2.1 从 GAN 到扩散：一次范式转移 45
2.2 加噪与去噪：从噪声里"雕刻"出一张画 47
2.3 潜空间：为什么模型要在"压缩的世界"里画画 48
2.4 U-Net 与 DiT：到底是谁在负责去噪 50
2.5 文本条件与无分类器引导：提示词到底怎么起作用 51
2.6 流匹配：为什么 2024 年后成了主流 53
2.7 参数语义：每个旋钮"调大调小"会发生什么 54
2.8 编辑工具箱：从重绘到条件控制 57
2.9 角色一致性：让同一个人出现在不同画面里 60
2.10 工具地图：2026 年 10 月，你能用上什么 62
2.11 关键概念表 67
2.12 案例教学：从文字分镜到六张关键帧 68
2.13 失败案例诊所：症状、成因与处方 71
2.14 动手实践：从文生图到关键帧的四个阶段 75
2.15 课程思政：用图像讲中国故事，也守好取材的边界 78

第 3 章 · 视频生成与编辑 83

3.1 视频的时空表征：多出来的那一维 83
3.2 三条技术路线：视频模型是怎么"长"出来的 86
3.3 时间一致性：为什么"误差"会自己长大 89
3.4 让画面自己"出声"：音视频联合生成 91
3.5 条件控制：你想让它怎么动，就怎么"喂"它 92
3.6 参数语义：为什么你的"中等运动"不是我的"中等运动" 99
3.7 工具地图：2026 年 10 月，你手里有什么牌 102
3.8 编辑与后期：把"抽卡"变成"作品" 105
3.9 现实约束：一个镜头要"抽"几次才够 107
3.10 失败案例诊所：把"翻车"翻译成"处方" 108
3.11 关键概念表：把这一章拧成一把钥匙 111
3.12 动手实践：把分镜"动"起来 113
3.13 真实与标识：让 AI 作品"说真话" 115

第 4 章 · 三维模型生成与编辑 120

4.1 三维数据在计算机里长什么样：表征家族 120
4.2 两条来路：重建与生成 120
4.3 让 AI"想"出三维：从 SDS 蒸馏到原生三维 126
4.4 纹理与多视角一致性：为什么"正面好看"不可信 129
4.5 拓扑、面数、UV 与 PBR：把模型"修成能用的样子" 132
4.6 导出、交接与检查清单：让资产"能出货" 135
4.7 工具地图：这一排名字，各管一段路 139
4.8 失败诊所：三类高发缺陷的完整病历 141
4.9 动手实践：给三个道具做"出厂检验" 143
4.10 关键概念表：本章词汇的"速查手册" 145
4.11 课程思政：文化遗产数字化复原的边界与真实性 146
4.12 本章产出与下一站 148

第 5 章 · 角色动画生成与编辑 148

5.1 从静止到会动：动画到底在动什么 151
5.2 骨骼：角色的"骨架"与运动链 152
5.3 蒙皮与权重：网格"听谁的" 154
5.4 关键帧与插值曲线：动画的"标点符号" 156
5.5 IK 与 FK：两种"让手动起来"的方式 158
5.6 动画十二原则：给眼睛配一把尺 159
5.7 运动条件与动作重定向：把动作"翻译"到另一个角色身上 161
5.8 AI 工具现状：绑定、文生动作与动作库 163
5.9 材质与缺陷修复：把角色"体检"合格再动起来 166
5.10 关键概念速查：本章的"词汇表" 170
5.11 动手实践：「举杯邀月」从绑定到交付 170
5.12 课程思政：AI 辅助，不等于放弃艺术标准 177

第 6 章 · 人体姿态与动作迁移 181

6.1 姿态：把一个人的动作变成一串坐标 181
6.2 三维姿态与人体模型：从"投影"回到"身体" 186
6.3 抖动、遮挡与平滑：把毛刺磨平 188
6.4 姿态条件控制：让 AI 按你指定的姿势作画 191
6.5 重定向：把动作交给另一个身体 194
6.6 "参考动作驱动角色"的现状：一条视频，一次模仿 196
6.7 失败案例诊所 198
6.8 案例教学：从舞蹈教室到屏幕 202
6.9 关键概念表 203
6.10 动手实践：从一段自拍到一个角色动作 204
6.11 课程思政：身体影像的使用边界与动作文化 208

第 7 章 · 游戏场景生成与编辑 212

7.1 场景的五块拼图：地形、建筑、道具、光照、氛围 212
7.2 场景好不好用：三个评价维度 214
7.3 把一句话变成一张清单：大语言模型与场景规格表 216
7.4 程序化生成：让规则替你"摆"场景 221
7.5 资产生产管线：每一类资产有它自己的"生产线" 224
7.6 落进引擎：导入流程、常见坑与性能预算 227
7.7 关键概念表 230
7.8 案例教学：三支团队，一条共同的流水线 232
7.9 失败案例诊所：三个真实的翻车现场 234
7.10 动手实践：搭出你自己的"夜江场景" 235
7.11 课程思政：文化准确性与团队劳动 241

第 8 章 · 智能体与交互媒体创作 244

8.1 智能体：四个零件，一个"会办事"的系统 244

8.2 ReAct：想一步、做一步、看结果 246
8.3 工具调用：模型怎么"伸手"到世界 247
8.4 MCP：给工具装一个"统一插座" 250
8.5 记忆与知识库：让它"记得住、查得到" 251
8.6 多智能体：分工、评审、辩论——以及一笔必须算清的账 254
8.7 验收与停止条件：为什么"让 AI 自己说完成"不可靠 256
8.8 零代码平台：把工作流"搭"出来 259
8.9 关键概念表 261
8.10 交互媒体：当智能体走进作品内部 263
8.11 案例教学：主线收官与三个可复用的样例 264
8.12 失败案例诊所 267
8.13 动手实践：从拆任务到公开验收 267
8.14 课程思政：判断在人，责任在人 272
8.15 回望全书：一条流水线的八个站点 273

附录 A–F • 随查手册：术语 • 工具 • 提示词 • 评分 • 合规 • 选学 276

附录 A 术语表：中英对照速查 276
附录 B 网页工具速查（2026 年 10 月版） 280
附录 C 提示词模板库（可直接复制） 281
附录 D 期末综合作品指南与验收标准 284
附录 E 合规与授权操作清单 286
附录 F 选学：最小 API 调用与本地工作流 287

导论

当生成变得廉价，创作还剩下什么

先看懂这门课的地图、工具与规矩，再出发。

学习目标

1. **知道：**这门课学什么、八章之间是什么关系、期末你要交出什么。
2. **理解：**生成式人工智能"是怎么学会创作"的——三次浪潮、拐点与规模定律的直觉。
3. **建立：**一套贯穿全书的"提出需求—生成—评估—人工修正—发布"工作方法。
4. **认同：**使用人工智能的三条合规底线与"人在回路"的创作立场。

开课前必读　建议 90 分钟　对应课程第 1 周导读

> 2026 年的一个普通周三，下沙，C417 机房。你可能已经用手机里的某个 App 修过图、写过作业、生成过一段配乐——甚至不知道自己用的就是"生成式人工智能"。这门课从一个朴素的问题开始：当机器的"生成"变得像打字一样廉价，"创作"这件事还剩下什么？本书立场鲜明：剩下的，是判断、是结构、是人的表达——而这三样，恰好都是可以学的。接下来 32 个学时，我们一边把原理讲透，一边带你把这八种媒介亲手做一遍。

0.1 为什么是现在：三个绕不开的事实

先看三个事实。它们解释了为什么 2026 年的一年级大学生，需要专门上一门"AI 创作"的通识课。

◆ 事实一：AI 能生成的，已经覆盖了整个媒体光谱

十年前，"AI 创作"基本等于"AI 作诗"——结果常常让人发笑。而今天，从一段文字、一句配音、一张海报、一段视频，到三维模型、角色动画、舞蹈动作、游戏场景，乃至能自己调用工具、分步骤完成任务的"智能体"，几乎每一种数字媒体形态都有了可用的生成工具。本书要带你把整条光谱走一遍——不是为了记住工具名字（它们每几个月就换一茬），而是为了理解**每一种媒介背后"机器在学习什么、控制什么"**这一层不变的东西。

◆ 事实二："能做出来"和"做得好"之间，隔着一整门手艺

你可能听过一种说法："现在只要说一句话，AI 就能给你任何东西。"这句话一半对、一半是幻觉。对的一半：工具确实把"从零到一"的门槛拉到了几乎为零。是幻觉的一半：生成结果的默认水平，永远只是"及格线"。同一句提示词，有人得到一张能上封面的画，有人得到六个手指的怪物——差别不在运气，而在于：提示词怎么组织、参数怎么调、哪一步该让 AI 做、哪一步必须人来做。这些"手艺"就是这门课的核心，也是期末作业与"随手玩"的分水岭。

◆ 事实三：这已经不是一个"兴趣"问题，而是一个"位置"问题

2025 年 8 月，国务院印发《关于深入实施"人工智能+"行动的意见》（国发〔2025〕11 号），明确提出到 2027 年新一代智能终端、智能体等应用普及率超过 70%，到 2030 年超过 90%。换句话说：你毕业时进入的任何行业，都会默认"会用 AI 创作"是一项基础办公技能——就像今天默认你会用搜索引擎和表格软件一样。这门课想让你在那个默认里，处于"会"的那一边，而且不是"会用某一个工具"，是"理解这一代技术"。

> 📌 **三个事实指向同一句话**
>
> 技术已经就位、工具已经平民化、政策已经写进国家行动——**真正稀缺的不再是"生成能力"，而是"指挥生成的能力"**：把模糊的想法变成清晰的需求（提示词与结构化表达），把生成的结果变成合格的作品（评估与人工修正），把作品变成负责任的表达（标识、授权与边界）。这门课教的，就是这三件事。

0.2 一张地图：八种媒介，一条流水线

本书的八章正文，不是八个彼此独立的"工具说明书"，而是一条流水线上的八个工位。请先记住这条流水线的样子：

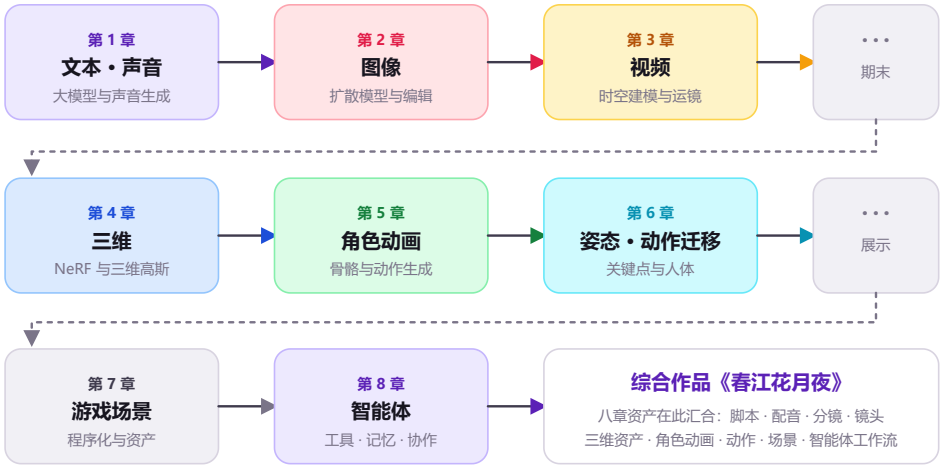


图 0-1　全书地图。八章 = 一条创作流水线的八个工位，从文本与声音出发，到智能体收束；每一章产出的资产都会成为后续章节的原料，最终汇聚成期末综合作品。

表 0-1　八章一览：每章你会亲手做出什么

章	主题	一句话内容	主线项目产出
第 1 章	人工智能与多模态基础 · 大语言模型与声音生成	机器怎么"读"文字、"听"声音、"说"人话：四族技术（大模型/语音识别/语音合成/音乐音效）与提示词的四要素	分镜脚本 + AI 配音音频
第 2 章	图像生成与编辑	扩散模型加噪去噪的直觉；文生图、局部重绘与条件控制	角色设定 + 4–6 张分镜关键帧
第 3 章	视频生成与编辑	时空注意力与时间一致性；运镜、首尾帧与剪辑流水线	4–6 段镜头视频
第 4 章	三维模型生成与编辑	三维表征七家族；SDS 蒸馏与原生三维生成；导出质检	楼台/扁舟等三维初稿 + 缺陷记录
第 5 章	角色动画生成与编辑	骨骼、蒙皮、关键帧；动作重定向与 AI 动画工具	诗人角色「举杯邀月」动画
第 6 章	人体姿态与动作迁移	2D/3D 姿态估计；遮挡与平滑；把真人动作搬到角色上	舞者动作提取与迁移
第 7 章	游戏场景生成与编辑	LLM 结构化需求 → 程序化布局 → 多模态资产 → 引擎落地	夜江场景 + 结构化资产清单
第 8 章	智能体与交互媒体创作	工具调用、记忆与多智能体分工；作品验收与公开展示	智能体工作流 + 期末作品

◆ 贯穿案例：为什么是《春江花月夜》

全书用一条主线把八章串起来：三个虚构的同学——小江（负责文案与策划）、小月（负责美术与视觉）、小舟（负责技术与流程）——要在期末交出一部约 2 分钟的古诗词数字短片，篇目定在张若虚的《春江花月夜》。

选它，有三个理由。第一，**它天生是"多模态"的**：全诗三十六句，几乎每一句都是一幅画、一个镜头——"春江潮水连海平"是空镜，"海上明月共潮生"是运镜，"江畔何人初见月"是独白，"谁家今夜扁舟子"是人物。第二，**它足够难**：写意、留白、抒情，恰好是生成工具最不擅长的部分，逼着你练"人工修正"的真

功夫。第三，**它足够中国**：这首诗被闻一多先生称为"诗中的诗，顶峰上的顶峰"，用最新的生成式技术表达最古典的意境，正是这门课"课程思政"的天然载体——技术是新的，文化根脉是深的。

你会看到，小江他们在每一章都会推进一格：第 1 章写脚本、配音；第 2 章画分镜；第 3 章出镜头；第 4 章捏出楼台与扁舟的三维模型……到第 8 章，这些零散的资产被一个多智能体工作流串成成片。从第 1 章开始，他们就正式出场。

图 0-2　贯穿案例的主视觉。接下来八章的所有技术，都服务于屏幕上这样一幅画面。这张图是「目标态」，也是全书「从提示词到作品」的具象化。

AI 生成

0.3 机器是怎么学会"创作"的：三次浪潮与一个拐点

要指挥 AI 创作，先要知道它是怎么"学会"的。这一节的篇幅不长，但它决定了你看待后面所有工具的眼光：**你不会再把 AI 当魔法，而是当成一个统计学意义上的"超级学徒"**。

◆ 三类任务：识别、生成、决策

人工智能在媒体领域干的事，大致分三类：

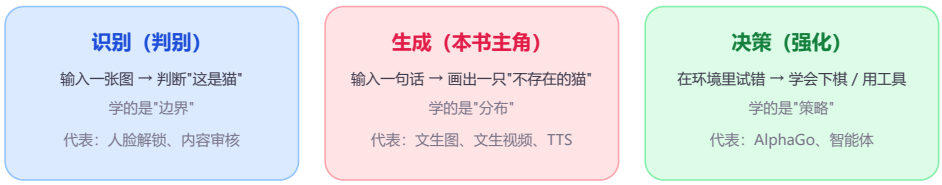


图 0-3　三类任务的分工。传统 AI 擅长"识别"，生成式 AI 的突破在于"生成"，而智能体把"决策"加了进来——三者叠加，才是"AI 辅助创作"的全貌。

识别（也叫判别）回答"这是什么"：人脸解锁、垃圾邮件、内容审核，都是在画一条边界。它不创造任何新东西。**生成**回答"它长什么样"：给它一句话，它画出一只从未存在的猫——它学的不是边界，而是数据的"分布"，也就是"什么像猫、什么不像"。**决策**回答"下一步做什么"：AlphaGo 落子、智能体决定调用哪个工具，都是在试错中学"策略"。本书的主角是第二类和第三类。

◆ 三次浪潮：从"规则"到"数据"再到"规模"

AI 七十年，走过三次大的浪潮。理解这三段历史，你就明白为什么 2022 年之后一切都变了：

图 0-4　三次浪潮与一个拐点。规则时代靠人写规则，学习时代靠数据学边界，生成时代靠规模学到"分布"；2022 年后的拐点，让生成第一次达到了"可用"的水平。

第一次浪潮（1956 起）·规则时代。研究者相信智能可以"手写"出来：把专家的知识编成一条条 if-else 规则。它在下棋这类规则明确的领域偶有亮点（1997 年"深蓝"战胜卡斯帕罗夫），但一旦遇到"识别一只猫"这种说不清规则的任务就立刻失灵——因为没有人能写出"猫的定义"。

第二次浪潮（2012 起）·学习时代。思路反过来：不写规则，让机器从海量数据里自己总结规律。2012 年，深度神经网络 AlexNet 在图像识别竞赛中大幅领先，深度学习从此引爆。这一代的机器学会了"识别"——认得猫、认得脸、认得你的语音。但它还不会"创作"。

第三次浪潮（2022 起）·生成时代。2017 年提出的 Transformer 架构（一种以"注意力"为核心的网络结构）改变了一切：它让模型可以吞下整个互联网的文本、图像与音视频。当数据、算力、架构三个条件同时成熟，"规模定律"（scaling law）开始显灵——模型越大、数据越多，能力不只是变好，而是不断"涌现"出新技能。2022 年，ChatGPT 与扩散模型先后公开，生成第一次达到了"普通人可用"的水平。此后的故事你在新闻里都见过：文生视频、三维生成、推理模型、智能体……一路到"人工智能+"写进国家行动。

∫ 深入一格 · "生成"到底在学什么　选学 · 数学直觉

把一亿张猫的图片想象成"图片空间"里的一亿个点。判别模型学的是在这片空间里画一条边界："边界这边算猫，那边不算。"而生成模型学的是**这些点的分布**——哪里密、哪里疏、什么形状。学到分布之后，模型就能在这片空间里"取样"：生成一只从未存在过、但看起来完全合理的猫。

$$\text{生成模型的全部工作：估计 } p(x) \text{ —— "什么样的 } x \text{ 才像真实数据"}$$

后文会看到：扩散模型把这件事拆成"一千步去噪"，语言模型把它拆成"下一个词元是什么"——本质都是对分布的逼近。

这个视角解释了本书后面几乎所有现象：为什么生成有随机性（取样本身就是随机的）、为什么"提示词"有效（它在引导取样的方向）、为什么会有幻觉（分布里没有的东西，模型会"编"出最近似的）、为什么调参数能改变风格（改变取样的搜索范围）。

0.4 能力边界：它擅长什么、又做不好什么

在动手之前，先建立一张"能力边界"的地图。这门课最浪费时间的事，就是拿 AI 不擅长的事去为难它，然后得出"AI 不过如此"的结论；同样常见的是拿它的强项当自己的功劳，结果期末"技术说明"写不出任何方法。看清边界，你才知道劲往哪儿使。

图 0-5　生成式 AI 的能力边界（本书写作时点）。左栏是"交给它"的清单，右栏是"必须人管"的清单——这张图会随你在各章的实践不断被验证。

把这张图记成两句话：**它擅长"在你已经想清楚之后高速生产"，不擅长"替你想清楚"；它可以给你十个答案，但不能替你承担一个责任。**前者是方法问题（提示词、参数、迭代），后者是伦理与法律问题（授权、标识、诚信）——恰好就是这本书的两条主线。后面每一章，我们都会在这张图上给具体的工具和技术"标坐标"：它落在左栏还是右栏，落在哪里，你就知道该怎么用它。

0.5 这本书怎么用：八段体例与四条轨道

本书每一章都按同一套"八段体例"写作。第一次翻开时花两分钟认识它，之后你会用得很顺：

图 0-6　每章的八段体例。原理、案例、实操三条线并行推进，思政元素嵌入其中——这是通识课教材"既要讲透、又不能用不上"的平衡术。

◆ 四条轨道：一眼识别正在读什么

全书内容用四种颜色的"轨道块"标记，方便你快速定位。它们在网页版和 PDF 里长得一样：

蓝 • 原理

机制、图解、路线图。慢读，这些是"内功"——工具换了，它们不变。

紫 • 案例

有背景、有做法、有结果、有复盘的真实感案例。看别人怎么趟坑。

青 • 实操

提示词卡、分步操作、参数对照。边读边做，读完全书你会攒下几十组可复用的参数记录。

红/金 • 边界

红色是"失败诊所"与合规提醒，金色是课程思政。别跳过——就业后最先用上的是这部分。

> 🧪 **网页版与 PDF 版的差别**
>
> 本书有**网页版**（可交互）与 **PDF 版**（便于打印与批注）两个形态，内容完全相同。差别有三个：网页版的**提示词卡**带"一键复制"按钮；**自测题**默认折叠、点击展开（PDF 版直接给出答案）；少数演示图在网页版里有动画（如扩散加噪过程），PDF 版呈现为动画的首帧——不影响理解。

> 🖼 **关于本书的插图**
>
> 本书插图分两类，来源都在图注里写明：**机制示意图**（流程图、结构图、参数曲线等）全部是手绘矢量图（SVG），保证标注精确、印刷锐利；**贯穿案例的视觉资产**（主视觉、分镜关键帧、角色设定、场景概念图，以及第 2 章的参数对照实测图）由 **AI 图像模型生成本地推理**产出，均在图注中标注 `AI 生成`——生成工具、提示词与参数记录见 附录 F 的说明。这么做有两个原因：其一，这门课的内容本身就是"用 AI 创作"，用真实的生成图像做示例，比任何示意图都更诚实；其二，本书自己在实践它教给你们的规矩——凡 AI 生成内容，必须如实标识。

◆ 学时分配：32 学时怎么花

本课程共 32 学时 = 课堂讲授 16 学时 + 课内实践 16 学时，横跨 11 教学周。八章各配 3–6 学时，每周的安排如下（与任课教师公布的授课提纲一致）：

表 0-2　教学周历与学时分配

周次	章节	学时	讲授内容要点	实践内容
第 1 周	第 1 章	3	AI 与生成式 AI 发展脉络、多模态数据与表征、伦理与合规框架	工具环境与账号准备；文本生成与提示工程初训；语音识别初体验
第 2 周	第 1 章	3	大语言模型架构、预训练与后训练；声音表征与条件生成	解说脚本创作与音频生成
第 3 周	第 2 章	3	扩散的加噪去噪、潜空间、U-Net 与 DiT、文本条件与 CFG、流匹配概览	文生图、图生图、局部重绘与条件控制参数对照
第 4 周	第 3 章	3	视频时空表征与注意力、时间一致性建模	文字/参考图/首尾帧条件下的视频生成、比较与剪辑
第 5 周	第 4 章	3	网格、NeRF、三维高斯等表征；SDS 蒸馏、生成与重建、纹理	文生三维、图生三维与模型导出检查
第 6 周	第 5 章	3	骨骼、关键帧、蒙皮、运动条件与动作重定向原理	模型材质与缺陷修复，角色动画生成与人工修正
第 7 周	第 6 章	3	2D/3D 姿态估计、关键点与置信度、遮挡与平滑	姿态提取或姿态条件生成与错误排查
第 8 周	第 6 章	3	姿态条件控制、时序平滑、动作迁移与三维角色驱动	姿态数据迁移至三维角色并修正动画
第 9 周	第 7 章	3	LLM 结构化需求理解、程序化生成约束、资产与场景评价	场景方案设计与资产生成
第 10 周	第 8 章	3	API 与工具调用、检索增强与记忆、多智能体分工与验收	场景校验与修订、智能体任务分解实训
第 11 周	第 8 章	2	综合创作工作流、单/多智能体对照	作品公开展示、互评与教师验收

0.6 期末你会交出什么：评分即路线图

这门课没有闭卷考试。期末是一次"公开展示 + 作品验收"，评分构成本身就是一张路线图——它明确告诉你"什么算学好"：

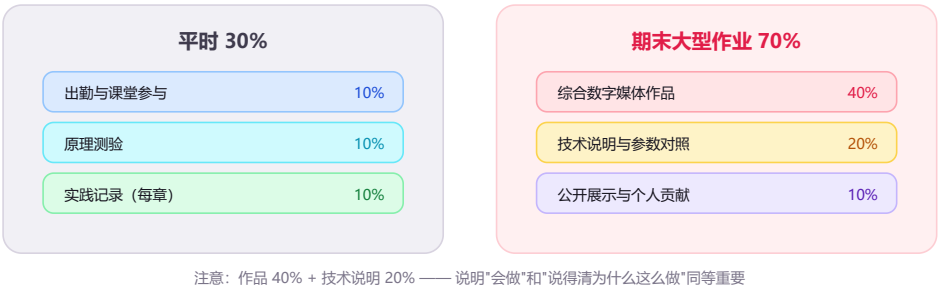


图 0-7　总成绩构成。"技术说明与参数对照"占 20%，这就是本书每章都逼你记录参数、做对照实验的原因——它是被计分的学习动作。

把 70% 的期末作业拆开看，其实只有三件事：**做出来**（综合作品，要求至少融合两类媒体）、**讲清楚**（技术说明：用了什么工具、什么参数、为什么，以及人工修改了什么）、**亮出来**（公开展示，并如实标注个人贡献）。你会发现，这三件事恰好对应本书反复强调的三句话：作品是"评估与修正"的终点，说明是"参数记录"的沉淀，展示是"合规与诚信"的考场。

> **🎯 一条贯穿全书的评分线索**
>
> 从第 1 章起，每章的"动手实践"都会提醒你保存三类东西：**成品**（图片/音频/视频/模型）、**生成参数或调用记录**（提示词、参数、种子）、**对照结果**（A/B 实验截图）。到期末，它们直接构成"技术说明与参数对照"的素材库——不要等期末再补记录，那补不回来。

0.7 工具、账号与三条合规底线

◆ 你不需要任何"高配设备"

本课程的实践以**网页版生成工具**为主：打开浏览器、注册账号、就能创作。第 1 周会集中做环境准备，附录 B 给了完整的工具速查表（入口、免费额度、替代品）。有独显电脑的同学可以选学附录 F 的"本地工作流"，但不是必需——课程不考"设备"，考"方法"。

◆ 三条合规底线（第一周就要背下来）

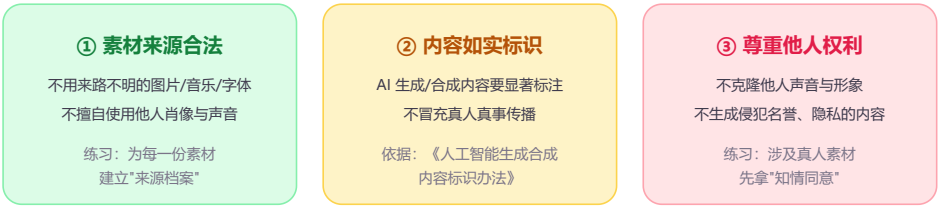


图 0-8　创作合规三原则。三条底线会在各章反复出现：素材、标识、他人权利——它们不是"额外负担"，而是作品能公开传播的前提。

这三条底线不是我们发明的，而是国家法规与平台规则共同划出的红线。第 1 章导论部分会展开讲法规原文与操作细节，附录 E 是随时可查的合规清单。此处先记住最实用的一句：凡是涉及真人（脸、声音、身体影像）的素材与生成，先问授权；凡是 AI 生成的内容，发布时如实标识。

> **🇨🇳 课程思政 · 做负责任的技术使用者**
>
> 同学们将来会用这门课的技术做毕设、做自媒体、进入广告/影视/游戏/电商行业。技术越强大，越考验使用者的价值判断：**用它放大善意还是制造混乱，用它传承文化还是制造噪声，用它服务真实还是伪造真实**——这些分岔口没有标准答案，但每个创作者都要做出选择。这门课的立场是：坚持科技向善，做负责任的创作者——对素材来源负责、对生成内容负责、对受众负责。这也是"人工智能+"时代中国青年技术素养的一部分。

0.8 人在回路：这门课真正想教的东西

最后一个问题，也是全书的立场：**AI 会取代创作者吗？**更准确的问法是：**AI 把创作者的工作改变了——那么，人应该站在流程的哪个位置？**

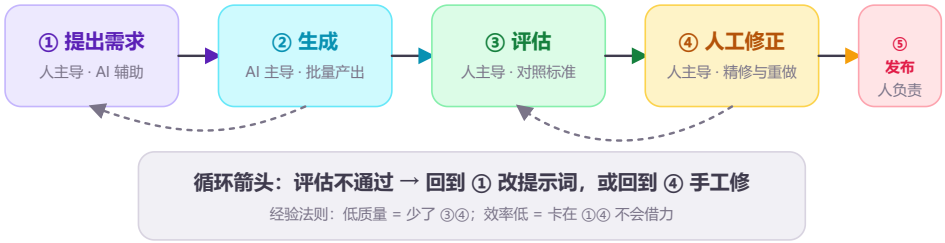


图 0-9　"五步工作法"中的人机分工。AI 在第 ② 步无可替代地强大，但第 ①③④⑤ 步——定义问题、判断好坏、亲手修正、承担责任——始终由人主导。本课程的能力目标就写在这张图里。

本书把这项工作方法总结为"五步工作法"：**提出需求 — 生成 — 评估 — 人工修正 — 发布。**你会发现八章的内容全都挂在这五步上：第 1 章教你"提出需求"的语言（提示词），第 2–7 章教你各媒介的"生成—评估—修正"的具体方法，第 8 章教你把整个流程交给"智能体"去规模化执行。而第 ⑤ 步"发布"的合规与责任，永远由人承担——这也解释了为什么本书把"课程思政"写进每一章而不是贴在最后：技术怎么用，始终是人的问题。

◆ 给同学的三条学习建议

1. **动手优先，但先想后做。**每读一节原理，先在脑中问"如果参数这样调会怎样"，再去工具里验证——带着假设做实验，比漫无目的地"抽卡"高效十倍。
2. **记录胜过记忆。**把每次像样的生成都存下"提示词 + 参数 + 结果 + 一句话复盘"。本书附录 C 给了模板，期末你会感谢自己。
3. **比较出真知。**所有"哪个更好"的问题，都用 A/B 对照实验回答：一次只改一个变量，看结果差异。这既是科研训练，也是本课程"参数对照"考核的底层逻辑。

◆ 把导论压缩成七句话

1. 生成式 AI 已经覆盖文本、声音、图像、视频、三维、动画、姿态、场景八种媒介，并且门槛降到"一句话"级别。
2. 但"能生成"不等于"能创作"——差距在提示词、参数、评估与人工修正这些手艺上，这正是本课教的东西。
3. AI 走过规则、学习、生成三次浪潮；2022 年是拐点，2025 年"人工智能+"写进国家行动，会 AI 创作正在成为基础技能。
4. 判别学边界，生成学分布，决策学策略——本书主角是"生成"，第 8 章加入"决策"（智能体）。
5. 全书八章是一条流水线，以《春江花月夜》数字短片为贯穿案例，每章产出一类资产。
6. 评分即路线图：作品 40% + 技术说明与参数对照 20% + 展示 10% ——记录、对照、说明，从第 1 章就要开始攒。
7. 三条合规底线：素材来源合法、生成内容如实标识、尊重他人肖像与声音权利；"五步工作法"中，人始终站在主导位。

◆ 自测

❓ 自测 1：为什么说"一句话生成"只是及格线？请用"分布取样"的视角解释。

答案：因为生成模型学的是数据的分布，给定提示词后它是在分布中"取样"。提示词越模糊，取样范围越大，结果越随机；要让结果稳定、精确地符合预期，需要把需求结构化（提示词）、约束取样方向（参数与参考条件），并对结果做评估和修正。所以工具把"及格线"变便宜了，但"优秀线"依然要靠人的手艺——提示词的精确度、参数的把控、对结果的判断力。

❓ 自测 2：本书的八章为什么是"流水线"而不是"八个独立工具课"？举一条资产传递的例子。

答案：因为真实创作中媒介是互相喂料的：第 2 章生成的分镜关键帧会作为第 3 章"图生视频"的输入图；第 4 章导出的三维模型是第 5 章做角色动画的对象，也是第 7 章场景里的道具；第 1 章写好的分镜脚本（先写"哪些镜头、每个镜头几秒"再做图、做视频）则是整条流水线的总谱。每一章的产出都是后一章的原料，最终在第 8 章被智能体工作流串成成片。"流水线"这个视角能让你提前知道每章要保存什么、交付什么。

❓ 自测 3："技术说明与参数对照"占期末成绩 20%，它考察的是什么能力？如何准备？

答案：考察的是"说得清为什么这么做"的能力——即把创作过程还原为可复现的实验：用了什么工具与模型、关键参数取了多少、如何做了 A/B 对照、人工修正了哪些部分、结论是什么。准备方式是每章实践后即时记录三类材料：成品、生成参数/调用记录、对照结果。这正是"提出需求—生成—评估—人工修正"工作方法的外化，也是未来职业中与团队协作、与客户沟通的基本功。

❓ 自测 4：用"识别—生成—决策"三类任务，给下列应用归类：① 短视频平台自动给视频打标签；② 用 AI 给纪录片配旁白；③ 智能体自动检查素材版权后生成发布清单。

答案：① 识别（判别）：判断"这个视频属于哪一类"，学的是边界；② 生成：从文本合成语音，学的是分布；③ 决策+工具调用：智能体需要规划"先查什么、再查什么"并调用外部工具，属于强化/智能体范畴（第 8 章）。现实中一个完整系统常常三类都用上。

❓ 自测 5：为什么本书把"课程思政"嵌入每一章，而不是放在书末集中讲？

答案：因为合规与伦理问题不是抽象的道德讨论，而是"下一秒就要做的操作"：选素材时要查来源、写提示词时要避开侵权内容、发布前要加 AI 标识、用真人影像前要拿授权。把这些判断放在它真正发生的技术语境里（比如第 2 章讲图像时讲素材来源合规、第 6 章讲姿态时讲身体影像授权），才可能真正变成行为习惯；集中放在书末只会变成一段没人读的说教。这也是本课程"把育人元素嵌入创作全过程"的设计意图。

◆ 拓展阅读

- 国务院《关于深入实施"人工智能+"行动的意见》（国发〔2025〕11 号）——建议通读一遍"六大行动"，看看哪些与你未来的行业相关。
- Ian Goodfellow 等《深度学习》（人民邮电出版社）——本课程指定教材，中文译本；本书第 1 章"深入一格"部分会用到它的第一章直觉。
- 皮埃罗·斯加鲁菲《人工智能通识课》（人民邮电出版社）——适合周末翻读的科普读本，了解 AI 的历史与人文面向。
- 国产大模型官方技术报告与文档（第 1 章会给出清单）——了解"中国方案"在生成式 AI 领域的最新进展。

第 1 章

人工智能与多模态基础 · 大语言模型与声音生成

机器怎么"读"文字、"说"人话——把地基打牢，后面八章才有地方站。

学习目标

1. **知识：**理解多模态数据的统一表示（词元、嵌入）、大语言模型的架构直觉与训练三阶段，并分清"语言与声音"上**四族不同技术**（大语言模型 / 语音识别 / 语音合成 / 音乐音效）各自在做什么。
2. **能力：**能完成生成式工具环境与账号准备；能用"四要素"写出可用的提示词并调控温度等参数；能创作解说脚本并用 AI 完成配音，也能用语音识别做字幕与口播剪辑。
3. **素质：**掌握幻觉核查的方法；了解中国人工智能治理法规框架，形成对声音权、表演者权的尊重意识。

对应第 1–2 周 · 6 学时　讲授 4 学时 · 实践 2 学时　主线推进：分镜脚本 + 配音

> "春江潮水连海平，海上明月共潮生。"小江想在短片开头用一段男声旁白念出这两句，配上潮水音效。她在对话框里敲下："请写一段《春江花月夜》的解说词，要有画面感。"十秒钟后，屏幕上出现了一段像模像样的文字——但读起来像旅游宣传册。问题出在哪？提示词没说清楚"给谁看、什么风格、多少字、分几个镜头"。这一章要解决的就是这件事：让你说的每一句话，都能变成机器听得懂的"指令"——顺带把机器听懂指令的原理、以及把文字变成声音的全过程讲透。

图 1-1　本章的四族技术地图。「语言与声音」是介质，不是技术；四族技术的模型、数据与失败模式各不相同。先认清它们的边界，再往下读，你才不会把「提示词的道理」误用在「音频参数」上。

1.1 多模态：机器眼中的世界，全都是数字

先立一个贯穿全书的观念：**在计算机内部，不存在"文字""画面""声音"这些东西，只有数字、向量和概率。**所谓"多模态"（multimodal），说的是把文本、图像、音频、视频、三维模型这些不同形态的信息（每种形态叫一个"模态"，modality），**翻译成同一种数字语言**，让一个模型能同时处理。

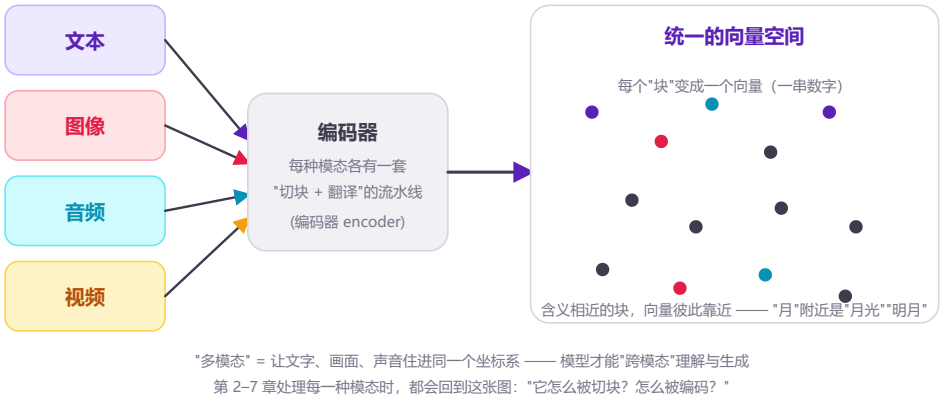


图 1-2　多模态的统一表示。不论哪种媒介，进入模型前都要经过"切块（token 化）+ 编码（embedding）"，变成向量空间里的点。这是全书最重要的心智模型。

表 1-1　五种模态：怎么切块、怎么编码

模态	切块方式（切成什么"词元"）	编码器做什么	典型应用
文本	分词器切成词元	把词元映射成向量，逐层融合上下文	对话、写作、翻译、代码
图像	切成小方块（patch，如 16×16 像素一格）	视觉编码器把每个小块变成向量，并学习块与块的空间关系	图像理解、文生图、图生文
音频	按时间切成短帧（如每 20 毫秒一帧），或经"神经音频编解码"压成离散码	把每帧的频率特征变成向量，学习时序上的"声音语法"	语音识别、语音合成、音乐生成
视频	切成"时空小块"：空间方向分块 + 时间方向分帧	同时编码"画面里有什么"和"前后帧怎么变"	视频理解、文生视频（第 3 章）
三维	表示形式多样：顶点序列、体素格、连续场采样的点	学习"形状的空间规律"（第 4 章详述）	三维生成、重建、场景建模

这张表值得多看两眼："切块 + 编码"是所有模态进入 AI 的统一入口。第 2 章的图像生成、第 3 章的视频生成，本质都是在处理"另一种词元序列"——你会发现它们的提示词写法、参数逻辑，与本章的文本提示一脉相承。这也是本书把"表征"放在第一章的原因：**看懂了这个入口，后面每一种媒介都只是换了"块"的形状。**

◆ 词元：机器眼中的"字"

模型不按"字"或"词"处理文本，而是按**词元**（token）：一种介于字和词之间的切块单位。切分规则由"分词器"（tokenizer）决定，大致可以这样理解：

- 英文里一个常见单词通常是 1 个词元，生僻词会被拆成两三个（如 `unbelievable` → `un` + `believ` + `able`）；
- 一个汉字通常是 1–2 个词元，取决于具体模型的分词表；
- 一个标点、一个空格也算词元。

为什么要知道这个？三个现实原因：第一，**几乎所有大模型都按词元计费和限长**——"你的输入不能超过多少词元"说的就是这个；第二，理解"词元≠字"，你就能理解模型的一些怪毛病，比如为什么它有时会把一个词的中间切开处理、为什么对数字的算术不敏感（数字也只是一个词元串）；第三，**不同厂商的分词方式不同**——同样的文本，在不同工具里的词元数可能相差不少，这就是"为什么同一段材料，A 工具装得下、B 工具装不下"的技术原因。凡是需要精确控制长度（如严格按字数生成的旁白）的场合，都要用目标工具实测一次，而不是凭感觉换算。

图 1-3　从句子到向量。切词元、再嵌入，是文本进入模型的第一步；第 8 章讲"检索增强生成"时，同一套向量表示会再次出场（用于检索相似知识）。

◆ 嵌入：把语义变成"坐标"

嵌入（embedding）指把每个词元映射成一串数字（向量）。关键在于：这串数字不是随便编的，而是在训练中"学"出来的——语义相近的词元，会被安排到向量空间里相近的位置。"月"和"明月"距离近，"月"和"冰箱"距离远。这一点也是后面所有技术的基础：只要能把某种媒介切块并嵌入，模型就能用同一套办法处理它——图像被切成小块（patch）嵌入、音频被切成小段嵌入、视频被切成时空小块嵌入。所谓"多模态大模型"，本质就是**一个统一的向量空间 + 多套编码器。**

∫ 深入一格 · 为什么"相近"可以用数字判断　选学 · 数学直觉

两个向量的"相近程度"，最常用的是**余弦相似度**：看两个向量方向的夹角。夹角越小，方向越一致，语义越近：

$$cos(A, B) = (A \cdot B) / (|A| \cdot |B|)$$

A · B 是点积（每个分量相乘再求和）；|A| 是向量长度。结果落在 −1 到 1：1 表示完全同向，0 表示无关，−1 表示相反。

你可以把它想成"指南针方向"：向量长度（说了多少）不重要，方向（说的是什么）才重要。"月"与"明月"的向量方向几乎一致，余弦相似度接近 1；"月"与"冰箱"的方向近乎垂直，相似度接近 0。第 8 章的"知识库检索"，就是拿你的问题去和知识库里所有片段的向量比相似度，取最接近的几条。

◆ 不止会说话的模型：视觉语言模型的预告

本章的主角是"语言"模型，但你实际用的工具里，多数已经能"看图"：上传一张分镜图问"这个构图有什么问题"，或者把照片里的场景描述成提示词——这是**视觉语言模型**（vision-language model）在起作用：在第 1.1 节那张图上，"图像"也被切块、编码，然后与文字向量放进同一个空间，于是"看图说话"变成了"在同一空间里匹配"。

为什么现在预告这件事？因为它正是第 2 章的起点：既然图像能被"翻译"成向量，那么反过来，从文字的向量"翻译"回图像就成了一条可行之路——第 2 章的扩散模型走的正是这条路。你现在建立的"文本 ↔ 向量 ↔ 图像"的心智模型，会在下一章立刻被用到；到第 3 章的"视频"、第 4 章的"三维"，只是再往向量空间里塞进"时间和空间"第三个、第四个维度。

◆ 潜空间：多模态的"公共休息室"

你已经看到，文本、图像、音频都被编码成向量。把这些向量所在的高维空间统称为**潜空间**（latent space，也译"隐空间"，本书统一用"潜空间"）。它有两个反直觉但极其重要的性质：

- **它比原始数据"小"得多。**一张 1024×1024 的彩图有 300 多万个像素值；编码后可能只有几百个数字。"潜"的意思就是"压缩后的、抽象的表示"——保留语义，丢掉冗余。压缩必然有损，这是后面所有"生成细节丢失"问题的总根源（第 2 章会细讲）。
- **在它里面可以"做算术"。**因为语义变成了方向，向量之间的运算往往对应语义的操作——"国王−男人+女人≈女王"是最著名的例子。第 2 章的"图生图"、第 6 章的"动作迁移"，本质上都是在潜空间里"沿着某个方向移动"。这张图先建立直觉，后文不断回访：

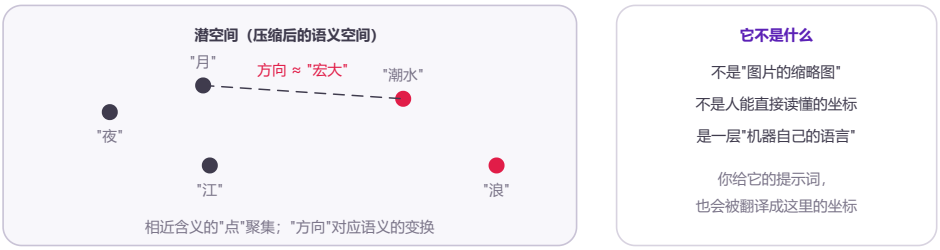


图 1-4　潜空间直觉。"语义相近 = 位置相近、语义变换 = 方向移动"——这个直觉将支撑第 2 章（图像）到第 6 章（动作）的全部原理讲解。

1.2 大语言模型：一台"预测下一个词"的机器

大语言模型（large language model，LLM）做的事，一句话可以说完：给定前面所有的词元，预测下一个词元最可能是什么。"春江潮水连海"后面接什么？模型会给所有候选词元一个概率："平"可能是 62%，"天"8%，"面"3%……然后按概率取样，选出"平"，再把它接到后面，继续预测下一个。整个"写作"过程，就是这件事重复几百上千次。

听起来平平无奇——但深想一层就可怕了：**要准确预测"下一个词"，模型必须在内部学会语法、常识、逻辑，甚至世界知识。**因为要接对"潮水连海平"，它得"知道"这是一句唐诗；要接对"如果 A>B 且 B>C，那么 A 和 C 的关系是"，它得会传递推理。这就是"预测下一个词"能涌现出翻译、写作、编程、推理能力的原因。

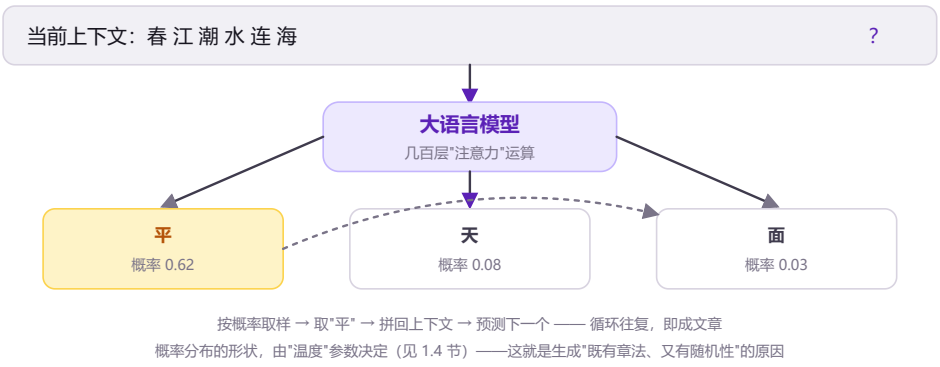


图 1-5　大语言模型的工作方式。每一步都输出"下一个词元的概率分布"，再按概率取样。所谓"写作"，是几百次取样串联的结果。

一次完整的生成：把"海上明月共潮生"拆开看

为了把"预测下一个词元"这个抽象说法变成可触摸的过程，我们把前几节的工具串起来，完整推演一次。假设提示词是"以《春江花月夜》的意境写一句七言诗，以'海上明月'开头"，模型的"思考"（其实是计算）大致是：

1. **读入与编码。**把你的提示切成词元、嵌入成向量。此刻模型"手上"是几百个向量。
2. **逐层计算注意力。**经过几十到上百层：每一层里，每个词都"看一圈"上下文（1.1 节的加权平均），不断融合信息。到最后一层，"下一个词"的候选分布已经蕴含了语法、意象搭配与格律倾向。
3. **取样第一个字。**在"共/同/连/随……"的候选里按概率取到"共"——注意，这一次取样是随机的，换一次可能取到"同"（这就是温度在起作用）。
4. **回填与循环。**"海上明月共"成为新的上下文，模型继续预测下一个字："潮"。如此往复——潮、生、。、然后停止。
5. **终止条件。**模型输出一个特殊的"结束词元"，或达到你设置的长度上限，生成结束。

五个步骤里藏着本课程后面每一章都会重逢的四个概念：**嵌入**（第 2 章图像的"块"）、**注意力**（第 3 章视频的"时空"）、**取样**（全部生成工具的参数语义）、**终止条件**（第 8 章智能体的"停止条件"）。大语言模型是这套机制的第一个实例，不是唯一的实例——把这五个步骤吃透，后面只是"词元换成了图像块、音频帧、三维点"。

◆ Transformer 与注意力：它到底怎么"读懂"上下文

支撑这一切的架构叫 **Transformer**，核心机制是**注意力**（attention）。用一句人话来概括：处理每个词时，先"看一圈"上下文里所有词，决定该重点关注谁。

看这个例子："小月把画架搬到窗边，因为**它**光线最好。"读到"它"时，模型要给上下文里的词分配注意力：对"光线"、"窗边"权重大，对"画架"权重小——于是它知道"它"指窗边，而不是画架。这个"权重分配"就是注意力。**权重怎么算出来的？靠训练**：模型从海量文本中学会了"什么词该关注什么词"。

图 1-6　注意力的直觉。处理一个词时，给上下文每个词分配权重，再按权重把上下文"加权平均"进当前词——语义的指代、修饰、因果，都在这个过程中被编码。

∫ 深入一格 · 为什么中文对 AI"稍微难一点"　选学 · 语言与工程

英文按空格切词，天然有"词"的单位；中文没有词间空格，"南京市长江大桥"到底切成"南京市/长江大桥"还是"南京/市长/江大桥"，需要模型自己判断。这使得中文分词（以及词元切分）比英文更容易出错，也解释了你在实践中会遇到的一些现象：

- **中文词元更"贵"**：一个字常占 1–2 个词元，同样意思的中文文本，词元数往往比英文对照更多——估算上下文窗口时要考虑到这一点。
- **逐字生成更像"手写"**：中文生成是逐词元（近似逐字）推进的，所以模型对"整个词"的把握不如英文稳，偶发"形近字""同音字"错误。
- **文化与典故是另一层门槛**：新模型的中文能力近年来提升显著，但涉及诗词、典故、地方文化的准确性仍需人工把关——这也是我们坚持"引文回原典核对"的原因。

选修提示：这部分内容与"中文信息处理"方向的研究一脉相承，感兴趣的同学可以从"分词与词元化的区别"这个关键词开始延伸阅读。

∫ 深入一格 · 注意力就是一次加权平均　选学 · 数学

每个词的向量会被投影成三个角色：**查询**（Query，我在找什么）、**键**（Key，我是什么）、**值**（Value，我能提供什么）。当前词的查询与所有词的键做点积，就得到"该关注谁"的分数；经 softmax 归一化成权重；再用权重对所有值加权求和：

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d}) \cdot V$$

QK^T：查询与每个键的相似度；$\sqrt{d}$ 是缩放因子（防止点积过大让 softmax 过于极端）；softmax 把分数变成和为 1 的权重；最后对 V 加权平均。

看，除了 softmax 和几个矩阵乘法，它就是"按相似度加权平均"。**整台 Transformer 的"智能"，就建立在这个朴素运算的成百上千层堆叠之上**——这也是本书反复出现的主题：生成式 AI 的复杂能力，往往来自简单机制的大规模组合。

◆ 训练三阶段：从"见过世面"到"会说话"再到"懂分寸"

一个大模型的能力不是一次训练出来的，而是分三个阶段"养"成的。理解这三段，你就明白为什么模型既博学又听话，也明白它的知识为什么有截止日期：

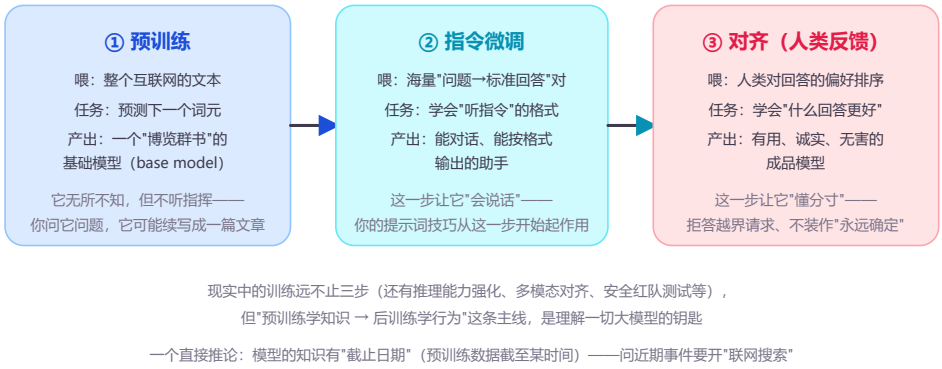


图 1-7　训练三阶段。"预训练学知识，后训练学行为"是理解大模型行为的钥匙：知识缺陷与幻觉来自第①段，听话程度与安全边界来自第②③段。

◆ 规模定律与"涌现"：为什么"大"本身就是方法

有一个反直觉的事实值得单独讲清楚：大模型的能力不是被"设计"出来的，而是被"规模"逼出来的。研究者发现，模型的能力随参数量、数据量、算力的增长呈现可预测的规律——这就是**规模定律**（scaling law）。更神奇的是"涌现"（emergence）：某些能力（如多步推理、指令跟随）在小模型上几乎为零，模型规模跨过某个门槛后却突然出现，像水在 0°C 突然结冰。

这解释了一系列现象：为什么业界拼命堆算力（因为"更大会更好"是可预测的）；为什么能力提升常常"突然"（涌现是相变式的）；也解释了你选工具时的直觉——**同一个家族的模型，大版本与小版本的能力差距可能是一道"能不能用"的门槛**，不要用轻量模型去硬扛复杂任务，也别用旗舰模型去做批量格式转换（浪费）。这条"按任务难度选模型档位"的意识，会贯穿后面每一章。

◆ 架构演化：模型在往哪儿长

2022 年以来，大模型虽多，演化方向却很清晰。你不需要记住型号，只需要记住四条趋势——它们决定了你未来能用工具做什么：

表 1-2　大模型的四条演化趋势（用"生活语言"解读）

趋势	是什么	对你的意义
更大：参数与数据持续扩张	模型参数量从千亿向万亿级迈进；"规模定律"让能力随规模继续提升	能力强弱有"档位"之分，重要任务用更强的模型，成本更高但值得
更省：混合专家（MoE）架构	把大模型拆成许多"专家子网络"，每次只激活相关的少数几个	同样能力下更快、更便宜——这是国产模型"高效率路线"的关键之一
更长：上下文窗口持续扩展	一次能读入的词元数从几千涨到百万级	可以整本小说、整个项目文档地喂给它——第 8 章"知识库"的物理基础
更全能：多模态与工具融合	同一个模型能看图片、听音频、调用搜索与绘图工具、进行"深度思考"	"一个助手干所有事"正在成为常态，跨工具协作是下一章起的主旋律

> **📌 中国方案：开源与效率路线**
>
> 这一轮大模型竞赛里，中国团队走出了两条有辨识度的路线：**一是高效架构**（如混合专家架构大幅降低推理成本），**二是开源生态**（众多国产模型开放权重，全球开发者可以自由部署与二次开发）。开源意味着：你不仅可以"用"模型，还可以"改造"模型——第 1 章实践环节与附录 F 会带你见识这条路。具体模型型号与版本更迭极快，附录 B 的工具速查表给出编写本书时（2026 年）的清单，使用前请以其官方文档为准。

◆ 上下文窗口：模型的"工作台"有多大

模型一次能"看到"的词元总量，叫**上下文窗口**（context window）。这个数字很关键，因为模型没有长期记忆——它每次回答时"知道"的一切，就是你这次会话里给它的内容（再加上训练时学进参数的知识）。窗口越大，"工作台"越大：可以一次丢进整本小说让它分析结构，可以把几十页的分镜文档发给它做统一校对。换算参考：一百万词元（1M）大致相当于几十万汉字——一本长篇小说的体量。

使用上下文窗口有四个实用认知：**第一，超长不等于好用**——内容太长时，模型对中间部分的注意力往往弱于开头和结尾（业界俗称"中间遗忘"或"迷失在中间"），重要指令建议放在开头或结尾，并显式重复一遍；**第二，"宣称窗口"往往大于"有效窗口"**：第三方长文本基准测试显示，不少模型能可靠利用的长度只有宣称值的六成左右——所以"塞满整个窗口"不如"精简后再喂"；**第三，窗口按词元算，不按字数算**，而且不同厂商的分词方式不同，同样一段中文在不同工具里的词元数可以差出五成，估算时要留余量；**第四，"没有记忆"是默认状态**——多数工具里，换一个对话框就"失忆"，要它延续上面的工作，得把关键结论重新贴给它（或使用平台的"记忆/项目"功能）。理解了这一点，你就明白第 8 章为什么必须讲"记忆机制"和"知识库"。

> ⚠️ **长上下文也"贵"**
>
> 把几十万字全塞进窗口，不仅可能触发"中间遗忘"，还会显著增加成本与等待时间（按词元计费时，输入越长账单越高；平台通常提供"缓存"机制降低重复内容的费用）。实用权衡：**几万字以内的材料，直接贴；再多的材料，先摘录要点，或改用第 8 章的"知识库检索"方案**——这也是"上下文工程"的一部分：给模型的信息，贵在"相关"而不在"多"。

1.3 推理模型：从"脱口而出"到"想好再说"

2024 年之后，主流大模型陆续多了一个开关：有的叫"深度思考"，有的叫"推理模式"（reasoning）。打开它，模型不再"脱口而出"，而是先写一段不给你看的草稿——反复推敲、自我检查、必要时回头重算——最后才给出答案。

这项能力的训练逻辑（用一个尽量准确的直觉）：**让模型在大量难题上练习"写出完整的解题过程"，并用结果对不对来奖励它。**久而久之，模型学会了"想得越充分、答对的概率越高"，于是自发地在难题上花更多"思考"。

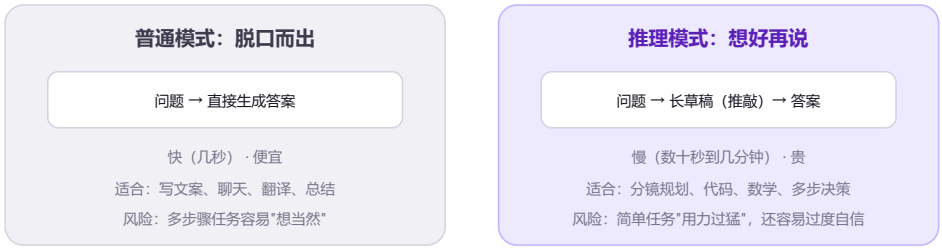


图 1-8　两种模式的分工。选模式的原则很简单：**任务需要"多步推理、自我检查"时开思考；需要"快和多"时关掉。**创作流程规划、参数调优方案适合前者，批量起标题、润色句子适合后者。

> 🎯 **创作场景怎么选**
>
> **开"深度思考"**：设计整部短片的分镜结构、排查"为什么生成的角色不一致"、规划智能体工作流、编写脚本代码。
> **关掉它**：起标题、写五版不同风格的台词、批量润色、整理会议纪要。
> 记忆口诀：要"想"，就开；要"快"，就关；不确定，先关——不满意再开。

> 🩺 **案例 1-C** § 用推理模式做一次分镜规划
>
> **背景：**小舟拿到小江的 8 镜头脚本，需要把它扩展到 2 分钟、约 16 个镜头的完整分镜，并规划每个镜头的"素材来源"（图片生成 / 视频生成 / 三维资产 / 现场实拍）。
>
> **做法：**他开启推理模式，把 8 镜头脚本、总时长、每章能用的工具清单一起贴进去，要求模型：① 先列出"哪些镜头可以复用同一张参考图"；② 再排出 16 镜头的顺序与时长；③ 最后标注每镜头的素材来源与风险（例如"泛舟镜头需要三维模型，工期长，建议放在第 4 章之后做"）。
>
> **结果：**模型给出的方案里，有一处它自己"回头修正"了——它最初把"落花"镜头安排给视频生成，随后在草稿里意识到"花瓣下落轨迹一致性要求高"，改为建议用图像序列 + 剪辑完成。
>
> **复盘：**这正是推理模式的价值：**它在输出前会"自我检查"，适合这种多约束的规划任务**；而同样的任务用普通模式，给出的排期经常忽略"资产必须先于镜头就绪"这条铁律。注意：推理模式的输出同样需要人工审核——它的"自我检查"只能发现逻辑矛盾，发现不了你不告诉它的现实约束（比如你的机房账号还没有三维工具）。

> ⚠️ **一个常见的"用力过猛"：不要对推理模型再加"请一步步思考"**
>
> 网上流传的提示词技巧里，"请一步一步思考"（思维链提示）很有名——**但它主要适用于普通模型。**对已经内置深度推理的模型，这句话是多余的，甚至可能干扰它自身的思考节奏。给推理模型的正确姿势是：把问题、约束和背景交代清楚，然后直接提问；需要它展示过程时，明确要求"给出关键步骤与依据"即可。另外注意产品层面的事实：不少 App 的"深度思考"开关**不会被记住**，每次新会话都要手动打开——这一点在机房里会让不少同学白等半天普通模式的结果。

◆ 把推理模式用好的四条经验

1. **给它"可检验的目标"。**推理模式擅长"有对错标准"的任务：算数、排期、逻辑检查。给它模糊目标（"写得有感觉"），它会用长篇思考换来平庸结果。
2. **要求它"先列假设，再动手"。**提示词里加一句"先列出你需要的假设与信息缺口"，能把它的思考引向"检查前提"——很多规划类错误的根源是默默接受了错误前提。
3. **让它自评。**"给出方案后，指出这个方案最可能失败的三个原因。"这一步几乎零成本，却常常直接产出下一轮的修改清单。
4. **别信"思考越长越好"。**简单任务配上推理模式，等于打车去买楼下的酱油——又慢又贵，还可能因过度分析把简单问题复杂化。任务难度与模式匹配，是判断力的体现。

> 🧪 **一个立刻可试的实验**
>
> 把同一个任务（如"为短片设计 5 条投票用的一句话简介"）分别交给普通模式与推理模式，记录：① 首字等待时间；② 结果可用率（5 条里有几条能直接用）；③ 你的判断。你多半会发现：推理模式给你"更完整的思考"但未必"更想要的文字"——这正是分流使用的依据，也是你写进"技术说明"的第一手证据。

1.4 提示工程：把"想要什么"说清楚的手艺

现在回到开头痛点：小江的提示词为什么只换来"旅游宣传册"？因为她只说了"要什么"（画面感），没说"给谁、多长、什么调、分几段"。**提示工程**（prompt engineering）就是研究"怎么说"的手艺。这一节给你一套能立刻用起来的框架。

◆ 四要素：角色 · 任务 · 约束 · 输出格式

一条"及格线以上"的提示词，通常包含四个要素。我们直接用小江的案例演示——把一句话扩写成一条能用的提示词：

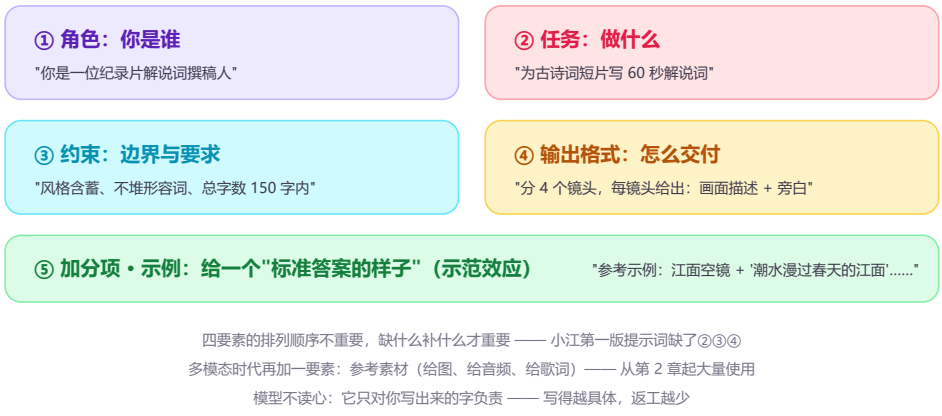


图 1-9　提示词四要素（+2）。角色定风格，任务定内容，约束定边界，输出格式定接口——"输出格式"是最容易被忽略、也最能提升可用性的一要素。

下面是改写后的完整提示词，你可以直接复制去试（提示词卡在网页版带一键复制按钮）：

✎ 提示词 · 解说词创作

你是一位纪录片解说词撰稿人，擅长"含蓄克制、以景写情"的撰稿风格。

任务：为一部 2 分钟的古诗词数字短片《春江花月夜》撰写开头 60 秒的解说词。

约束：
- 全片围绕张若虚《春江花月夜》前八句的意境，不要展开全诗；
- 不堆砌形容词，不出现"震撼""绝美"这类词；
- 总字数 **150 字**，句子以短句为主，便于配音朗读；
- 不虚构历史事实。

输出格式：分为 **4 个镜头**，每个镜头输出两行——
画面：（一行中文画面描述）
旁白：（一行中文解说）

参考示例：
画面：夜江全景，一轮明月自潮水中升起
旁白：潮水漫过春天的江面，月亮从水底升了起来。

模式 普通模式　温度 0.8　用到的要素 角色 · 任务 · 约束 · 格式 · 示例

再看一张"分镜脚本"的提示词卡——它演示了"把大任务拆成有格式的小任务"的思路（与上面的解说词提示词是配套的两步）：

✎ 提示词 · 分镜脚本（在解说词之后使用）

以下是我们已定稿的开头 60 秒解说词与风格说明。
【风格说明】含蓄克制、以景写情；画面留白多、人物少；色调由冷转暖。

【定稿解说词】
（此处粘贴上一步定稿的四段旁白）

任务：把它拆解为分镜头脚本表。

约束：
- **8–10 个镜头**，总时长 60 秒，单镜头 4–10 秒；
- 每个镜头标注：景别（远/全/中/近/特）、机位与运动（固定/推/移/摇）、画面主体；
- 为每个镜头写一条"可执行的 AI 生成提示词草稿"（中文，≤40 字）；
- 标注该镜头的可能风险（如"水面反光一致性难，建议考虑用视频生成而非静图"）。

输出格式：Markdown 表格，列为：镜头号 | 时长 | 景别 | 机位运动 | 画面主体 | 生成提示词草稿 | 风险备注

模式 深度思考　温度 默认　关键 表格列名 = 输出接口

注意最后一行"风险备注"——这是本书反复使用的技巧：在提示词里给它留一列"说实话的位置"。让模型评估自己方案的风险，比让它假装一切顺利有用得多：你会因此提前知道"水面反光"这类老大难在哪一镜等着你（第 3 章会验收这一点）。

◆ 参数：温度与取样

对话框里能调的参数不多，但每个都很关键。最重要的是**温度**（temperature）：

表 1-3　温度与取样参数速查

参数	它在做什么	调低（0–0.4）	调高（0.8–1.2）
温度 temperature	压平或拉尖"下一个词元"的概率分布——决定取样时的随机程度	输出稳定、保守、几乎可复现；适合事实问答、格式转换、代码	输出多样、有惊喜；适合起标题、找创意、写多版方案
核取样 top-p	只在"累计概率达到 p"的那一小撮候选词里取样，砍掉长尾	通常保持默认（多为 0.9–0.95）；与温度配合使用。一般只调温度、不动 top-p，除非工具明确要求	
输出长度上限	限制单次生成的最大词元数	写脚本、做规划时要留足；短任务设小可省钱。注意：截断不是"没话说"，而是被"喊停"	
随机种子 seed	固定取样的随机起点	想复现某次满意的结果：记下种子，同参数重跑。多数对话框工具不暴露它，第 2 章的绘图工具普遍支持	

温度 = 0.2 的实测表现

同一句提示词问三次，三次答案几乎一字不差；用词保守、结构端正，适合需要"标准答案"的任务。

但起草创意文案时会显得"温吞"，像不愿担责的实习生。

温度 = 1.0 的实测表现

三次答案三种风格，偶有惊艳的比喻，也偶有跑偏的句子。

适合"生成十个备选，人工挑一个"的工作流——记住：**高温度不是让 AI 替你决定，而是让它给你更多候选。**

四种常用提示模式：从"问一句"到"教一遍"

除了四要素，实践中有四种成熟的提示"套路"。它们按"你给模型的引导量"从小到大排列，引导越多，结果越稳定可控：

表 1-4　四种提示模式对照

模式	怎么做	适用	代价
零样本 zero-shot	直接描述任务，不给例子	日常问答、简单改写	结果不稳定，风格靠运气
少样本 few-shot	给 1–3 个"输入→输出"示例，让模型照猫画虎	格式要求严格的任务（如"每镜头两行"）	提示词变长；示例质量差会带偏结果
角色扮演	指定专家身份与口吻（"你是纪录片撰稿人"）	风格化写作、评审、模拟对话	不改变能力上限，只改变"说话方式"
逐步推导 chain-of-thought	要求"先列要点、再成文"或"一步一步分析"	多步骤任务、方案规划、数据核算	输出更长；简单任务反而变啰嗦

一个组合起来的例子（少样本 + 逐步推导，用于批量生成镜头文案）："请按以下格式为剩余 12 个镜头生成文案。**先**在草稿里分析每个镜头的情绪关键词，**再**输出成文。示例——镜头 1：情绪=辽阔→ 文案=……；镜头 2：情绪=孤独→ 文案=……"。"先分析、再成文"这个简单要求，常常能让文案质量明显提升，因为它把模型的"草稿空间"打开了——这也是 1.3 节推理模式的"平民版"用法。

迭代：没有人一次就写出好提示词

1. **先要一个"能看"的版本。**不要憋大招，用四要素写第一版，马上跑。
2. **指出不满意的具体点，让它改。**不要说"写得不好"，要说"第三句太浮夸，改成白描"——把"评估"变成下一轮提示词。
3. **一次只改一个变量。**同时改风格和字数，你就不知道是哪一项起的作用——这与后面各章的"参数对照实验"是同一套方法论。
4. **满意后，沉淀成模板。**把定稿提示词存进附录 C 那样的模板库，下次换主题直接替换占位符。

◆ 从"提示词"到"上下文工程"：把该给的信息给全

提示词的下一层功夫，是安排"**模型能看到的一切**"——不只是你最后敲的那句话，还包括：系统设定、历史对话、你粘贴的参考文档、上传的图片、工具返回的搜索结果。这个更大范围的功课叫**上下文工程**（context engineering）。它和提示工程的关系，好比"写一张便条"与"布置一张工作台"。

一个立刻能用的上下文工程实践是**把参考资料放进会话**，而不是指望模型"记得"：写《春江花月夜》的解说词时，把全诗原文、前几版定稿、风格要求列表一并贴进去，再让对方动笔。你会发现输出质量的提升，往往比"把提示词改得更花哨"大得多。另一个实践是**给上下文分段**：用标题（"【任务】【参考】【约束】【示例】"）把不同性质的内容分开——模型和人一样，读结构化的材料比读一大段"流水账"更不容易漏要点。

> 🧪 **系统提示词与多轮对话**
>
> 工具里的"系统提示词"（system prompt，有的界面叫"设定/角色设定"）是比用户输入优先级更高的世界设定：写一次，本轮所有对话都生效——适合定义"你是一位宋韵美学的视觉顾问"。而多轮对话的价值在于**上下文累积**：你可以把它已经给出的好句子"钉"住（"保持第二段的水准，重写第一段"），也可以用"我们刚才那版再来一稿，只改结尾"来精准迭代。对话不是聊天，是带着评审意见的返工。

> 🩺 **失败诊所 F1-1** "既要又要"的提示词
>
> **症状：**小月给模型下指令："把这段解说词改得简洁一些，最好再丰富一点，保持原味的同时大胆创新，字数不要变。" 结果模型输出了一版莫名其妙的文字，她连试了五次都不满意。
>
> **成因：**提示词内部包含互相矛盾的要求（简洁 vs 丰富、保持 vs 创新），模型只能在矛盾约束间随机"折中"——每次折中的方式不同，于是结果忽右忽左。这不是模型"笨"，而是它忠实执行了一条无法同时满足的指令。
>
> **处方：**①把矛盾拆开，一次只要求一个方向（先出"更简洁"版，再从简洁版出发要求"局部增添细节"）；②把模糊词量化（"简洁"→"每句 ≤20 字"；"大胆创新"→"把第三句的比喻换成与'潮水'无关的意象"）；③若两个目标确实都重要，就让它**分别生成两版**，人工融合——这正是"AI 给候选、人来拍板"的工作方式。

1.5 幻觉：模型什么时候会"一本正经地胡说"

小江请模型"引用一句古人描写春江的诗"，模型给出了一句查无此人的"名句"——平仄工整、气象宏大、完全不存在。这就是**幻觉**（hallucination）：模型生成了一段流畅、自信、但不符合事实的内容。

◆ 为什么会幻觉：三个原因

- **它在"取样"，不在"查证"。**回到 1.2 节的图：模型的每个词都是从概率分布里取的。"看起来最像下一句"的答案，不等于"真实存在"的答案——尤其当你的问题预设了一个错误的框架（"请引用×××的那句诗"），它会顺着你的框架编。
- **训练目标是"像"，不是"对"。**预训练奖励的是"预测得像人类文本"，后训练才会部分纠正"编造"——但纠正不彻底。
- **知识有截止日期。**预训练数据有截止时间，之后的事件它不知道；不知道时，一些模型宁愿"编一个像样的"也不说"不知道"。

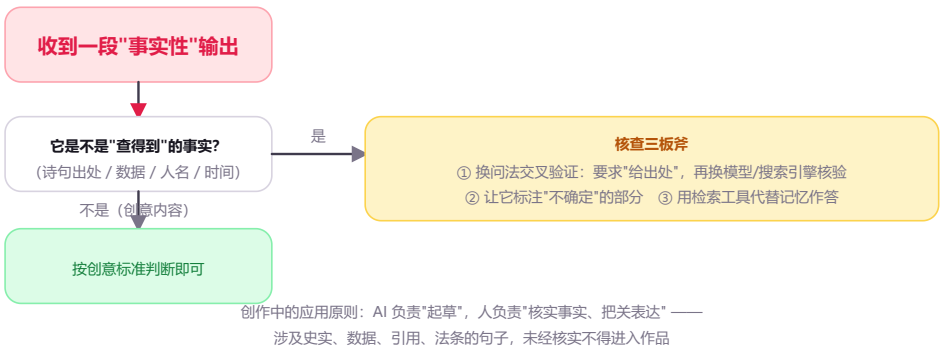


图 1-10　幻觉核查流程。先分类：这句话是"查得到的事实"还是"创意表达"？前者必须核查，后者按审美判断。这套"分类—核查"习惯，是使用生成式 AI 的基本素养。

⚠️ **三类高危幻觉（创作中最常踩）**

① **编造引用与出处**——"这句话出自《文心雕龙》"（其实不是）；② **编造细节**——把唐代的器物写成宋代的；③ **过度自信的数**——"这幅画拍卖了4.3 亿人民币"（听上去具体得可疑）。**凡是要写进解说词、字幕、技术说明的事实，都要二次核实**；无法核实的，删掉或模糊化。课程期末的"技术说明"里，如实说明"哪些内容经过核实、用什么方式核实"，会成为加分项。

◆ 三次"实测"告诉你幻觉有多普遍

幻觉不是理论担忧，而是被反复实测的问题。三个可以引用的公开案例：

- **"315 晚会"测试（2026 年 3 月，据上海辟谣平台等报道）**：向数个国产大模型提问"2026 年央视 3·15 晚会曝光了哪些品牌"，**仅 1 个模型回答正确**：有的把往年的曝光案例混入今年，有的断言"晚会尚未举办"，有的给出的报道链接实际打不开——错误还被后续引用，形成"错误链"。
- **中文史料引用基准的横评**：一项针对中文历史出处能力的研究测试中，全部参测模型共出现上百处"纪年不可能"的臆造来源（例如断言某书收录某事，而该书的成书年代根本不可能包含它）——说明"编造出处"是跨模型的结构性问题。
- **"过度顺从"的实验（Anthropic 团队论文，arXiv:2310.13548）**：研究发现，用人类反馈训练的模型会表现出"sycophancy"倾向——**当回答与用户观点一致时更容易被人和评分模型偏爱**；有时甚至为讨好用户而牺牲真实性。这意味着一个隐蔽的风险：**你说错了，它不但不纠正，还会顺着你编**。

🧪 **课堂实验：让模型"两边都同意"**

把同一个问题用两种立场各问一遍——比如"我认为《春江花月夜》重在写景，你怎么看？"再问"我认为《春江花月夜》重在抒情，你怎么看？"。观察模型是否两边都"你说得对、确实如此"。**这是全课程最便宜、最好复现的幻觉实验**：当它同时赞同两个矛盾观点时，你就亲眼看穿了"顺从"机制。防身用法：提问时不要"带货"——先中立地陈述任务，再让它给分析；要求它"如果你认为我的前提有误，请直接指出"。

◆ 核查的实操工具箱

知道"要核查"和"会核查"是两回事。下面是四种在你手边就能用的核查手段，按成本从低到高排列：

表 1-5　四种核查手段与适用场景

手段	怎么做	最适合核查	局限
① 要求给出处	直接问："这句话的出处是什么？请给出具体篇名/链接；如果不确定，明确说不确定。"	引文、数据、法条	模型可能"补"一个看似合理的假出处——出处本身也要核
② 换问法交叉	换个角度再问一次（"刚才说 X，那与 X 相反的说法有哪些？"），或换一个模型问同一问题	事实性判断、常识	两个模型可能错在同一处（同源数据缺陷）
③ 联网检索	打开工具的联网搜索，或自己去权威站点搜关键词	时效性信息、公开数据、新闻	搜索结果本身也有低质来源，需辨源
④ 权威源比对	回原典：诗句回诗集，法条回官方法规库，数据回官方统计或论文	要写进作品的"硬事实"	最慢——所以只对"要发布的关键事实"使用

使用的经验法则：**用①做普查（让模型自己标记不确定处），用②③做抽查，用④守住"要公开发布的关键事实"这道闸门。**别忘了把核查过程记进创作档案——期末"技术说明"里"本节涉及的三处史实分别经××方式核实"的写法，比任何形容词都有说服力。

◆ 幻觉的另一面：它是缺陷，也是"联想"的来源

有必要说清一件事：幻觉和创造力，在机制上是同一个东西的两个侧面。模型"编"出一个不存在的诗句，靠的是它把学过的意象、格律、修辞自由组合——同一套"自由组合"，换一个使用场景就是创意。所以我们的态度不是"消灭幻觉"，而是**分流**：

- **事实性任务**（写解说词里的史实、技术说明里的数据、引用）：以"零幻觉"为目标——核查、溯源、拿不准就不写。
- **创意性任务**（起名、比喻、分镜想象、"十个版本的片头语"）：欢迎"幻觉"——越出乎意料越好，人负责从候选里挑出那个"妙"的。

分流判断的标准很简单：**这句话如果错了，会不会有人受伤害？**会——进核查流程；不会——放手让它"胡思乱想"。第 2 章起你会发现，这个判断标准适用于所有媒介的生成结果。

◆ 怎么评估一段 AI 输出：四维打分法

课程期末的"技术说明与参数对照"要求你说明"为什么这一版更好"。说"感觉更好"不算数——你需要一套可复述的评估语言。下面这张表是本书推荐的"四维打分法"，后面每一章的实践都沿用它（各维度按具体媒介微调）：

表 1-6　输出质量四维评估法（文本示例）

维度	问什么	低分表现	高分表现
忠实度	是否符合需求与约束？涉及事实是否可靠？	跑题、违反约束、事实存疑	扣题精准，事实可溯源
可用性	能否直接进入下一步工作流？	还得大改格式、重写结构	格式即插即用，稍作润色即可交付
表现力	在审美维度上是否动人？	平庸、套话、模板味	有辨识度，有一两处"被击中"的句子
可控性	同一提示词再跑一次，质量稳定吗？	每次结果差异巨大	稳定在可接受区间——说明方法可靠，不是撞运气

用法：给每个维度打 1–5 分并写一句理由——**理由比分数重要**，因为它就是下一轮提示词的修改清单。"忠实度 2 分：把镜头 5 和 6 的内容写重了"——你看，反馈已经写成提示词了。

◆ 贯穿案例•第一步：小江的脚本工作流

> 🩺 **案例 1-A** § 从"旅游宣传册"到分镜脚本
>
> **背景：**小江负责《春江花月夜》短片（约 2 分钟）的脚本，第一期要交"分镜脚本 + 解说词"，对应教材第 1 章的产出。
>
> **做法：**她按四要素重写了提示词（见上一节的提示词卡），并用"一次只改一个变量"的办法迭代了四轮：第一轮调风格（从"大气磅礴"改"含蓄克制"）；第二轮调节奏（每镜头旁白压到 20 字内）；第三轮补约束（不许出现现代词）；第四轮把它写好的 4 个镜头作为示例，让模型扩展到 8 个镜头——把"示例"反向用：让模型自己产出的好结果当下一轮的示范。
>
> **结果：**拿到 8 镜头脚本初稿。她逐句核对了两处引文（发现模型把"滟滟随波千万里"错写成了"艳艳随波"，改回），并请模型为每镜头标注了"建议时长"。
>
> **复盘：**①第一版提示词缺三要素，返工时间反而更长——"想清楚再写"比"边写边想"快；②模型不知道"40 秒和 3 分钟的区别"意味着什么节奏，节奏必须人定；③引文必须回原诗核对——生成式 AI 的"文化常识"并不可靠。

1.6 语音识别（STT）：把声音变成文字

小江给短片做完了配音，下一步是字幕。她本能地打开剪辑软件想手打——六十秒的旁白，四段文字，逐句听、逐句敲，至少要半小时，还容易打错。小舟拦住她："你先看看这个。"他把音频拖进一个网页工具，十秒钟后，带时间戳的四行字幕整整齐齐地列在那里。

这就是**语音识别**（speech-to-text，STT，也译"语音转文字"、俗称"听写"）：把一段声音变成对应的文字。它是本章四族技术里的第二族，也是最被低估的一族——因为它的产物太"朴素"（不过是一段文字），以至于很多人忘了：没有它，就没有自动字幕、没有口播剪辑、没有可检索的播客文稿。

◆ 一条逆向的流水线：从波形回到文字

上一节开头我们说，声音是"机器记录的形态"（波形）与"接近听觉的形态"（频谱）的结合。STT 要做的，是从前者出发，倒推出后者背后"说的到底是什么"。现代系统的链路大致是三步：

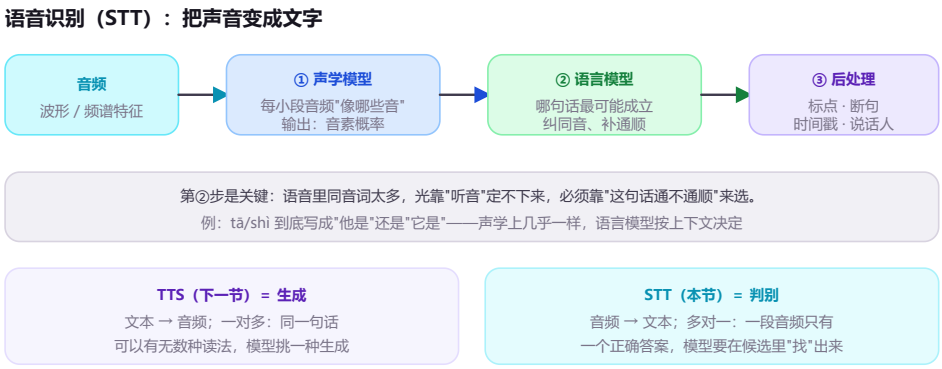


图 1-11　语音识别的三步链路，以及它与 TTS 的分工。同样处理"文字与声音"，STT 是**判别**（找最可能的文本），TTS 是**生成**（造一段像样的语音）——这正是第 1 章开头"识别 / 生成 / 决策"三类任务在声音上的具体化身。

◆ 创作者的四个用法：字幕只是最简单的一个

多数同学对 STT 的印象止于"自动字幕"，但它在创作流程里至少有四个位置，而且后三个比字幕值钱：

表 1-7　STT 在创作流程中的四个用法

用法	怎么做	省下的时间
① 自动字幕	识别音频 → 导出带时间戳的文本 → 导入剪辑软件生成字幕轨	手打字幕约占成片工时的两三成
② 口播剪辑	识别后得到"文字+时间轴"，**像改文章一样改视频**：删掉哪句话，就删掉对应片段	把"听音频找位置"变成"看文字点删除"
③ 素材检索	把几十段素材全部转写，用关键词搜出"说过某某"的那一段	素材越多收益越大——这是创作者最缺的能力
④ 文稿与笔记	访谈、会议、课程录音直接转成可编辑文稿	整理类工作的成本近似归零

注意②：它把视频剪辑的交互方式整个换掉了。当视频有了"文字层"，剪辑就从"看画面、找位置"变成了"读稿、改稿"——这是近几年内容行业最实在的效率变化之一，值得你在第 3 章的剪辑环节亲手试一次。

◆ 调好它：几个必须认识的参数

表 1-8　语音识别的常用参数

参数 / 功能	作用	什么时候必须动它
语言 / 语种	指定或自动检测要识别的语言	中英混说时，指定"中文为主"通常比自动检测更稳
热词表 / 自定义词	提前告诉模型"这些词很重要，别写错"	人名、地名、专业术语、品牌名——**效果最立竿见影的一项**
标点与断句	语音本身没有标点，标点是模型"加"上去的	需要做字幕时必开；写文稿时更要注意校对句子边界
时间戳粒度	按句输出时间戳，还是按词输出	做普通字幕用整句；做"卡拉 OK 式逐字高亮"才需要词级
说话人分离 speaker diarization	区分"谁在什么时候说"，输出带说话人标签的对话稿	访谈、多人播客、圆桌讨论
词错率 WER	识别的行业评估指标：错词数 ÷ 总词数	评估工具时看它；**但中文的同音字错误比这个词错率更值得你逐句核对**

🩺 **案例 1-1** § 小江的字幕流水线：十分钟做完六十秒

背景：《春江花月夜》前 60 秒的旁白做完了，要配字幕。

做法：① 把配音音频拖进识别工具，指定语言为中文；② 把"春江花月夜""滟滪""汀上白沙"等词加进**热词表**——这一步是她从小舟那里学来的；③ 识别得到四行带时间戳的文本，人工逐句核对（重点看同音字和人名）；④ 把结果导入剪辑软件，自动生成字幕轨，再微调两处换行位置。

结果：全程约十分钟，其中人工时间主要花在第③步的核对上。她原本打算花半小时手打。

复盘：三件事值得记住：①**热词表是投入产出比最高的一步**——几十秒的准备工作，省掉后面反复返工；②**核对不可省**——识别模型不认识"滟滪"，第一次输出的是"艳艳"，而这两个字在诗里意思完全不同；③字幕做完，视频就多了一条"文字层"，后面的剪辑可以直接按文字操作。

🩺 **失败诊所 F1-2** "治学"变成"自血"：同音字的连环坑

症状：一位同学把课程录音转成文稿交作业，老师读了两行就皱起眉头：'治学严谨'被识别成'自血严谨'，'期中'写成了'期终'，一位老师的名字更是每出现一次换一种写法。

成因：三件事叠加。①**同音字是中文识别的固有难点**：声学上"治学/自血"几乎无法区分，全靠语言模型按上下文选，而上下文越不常见，选错的概率越高；②**专业词汇与专有名词在训练数据里稀少**，模型更倾向于输出常见词；③**没有提供任何先验信息**——没有热词表，也没有告诉它这是哪门课的录音。

处方：①转写前先建热词表（人名、课程名、专业术语一次性喂进去）；②转写后**逐句核对**，重点盯三类词：同音字、专有名词、数字；③对要公开发布的内容（比如作品字幕），把"核对"当作流程的固定一环而不是可选项。识别工具给你的是"草稿"，不是"定稿"——和本章前几节讲的生成式工具一样，它的输出必须过人的判断这一关。

∫ **深入一格 · 为什么"听写"比"说话"难** 选学 · 信息论视角

初看之下，识别应该比生成容易：识别只要"找"出正确的答案，生成却要"造"出一个新东西。但工程上恰好相反——**高质量的语音识别往往比同等质量的语音合成更难做。**原因藏在两个方向的"映射结构"里：

- **生成是一对多**：念"海上明月共潮生"这句话，可以有无数种语速、音色、停顿的组合，只要听起来像样就行。模型的容错空间很大。
- **识别是多对一**：一段音频只对应一个正确的句子，但声学证据本身高度歧义——同一个发音可能对应几十个合法的词序列，模型必须在巨大的候选空间里"找到"唯一正确的那一个。

用一个信息论的说法：**生成是在输出端做选择（怎么表达），识别是在输入端消歧（听到的到底是什么）。**前者的不确定性可以"发挥"，后者的不确定性必须"消除"。

这个认识有直接的实用推论：给识别模型提供先验信息（热词表、领域提示、上下文），等于帮它把候选空间砍掉一大半——这与"给生成模型写清楚提示词"是同一件事的两面：**凡是让模型"少猜"的手段，都值得用。**

1.7 语音合成（TTS）与语音克隆：把文字变成人声

现在把方向倒过来：给定一段文字，要它"说话"。这是本章四族技术里的第三族——**语音合成**（text-to-speech，TTS）。它和上一节是一对镜像：STT 把声音变成文字（判别），TTS 把文字变成声音（生成）。而这一节还要讲一个 TTS 的"加强版"：**语音克隆**——不只让它说话，还要让它**用指定的音色**说话。

◆ 先补一块地基：声音是什么

要理解 TTS 的三段管线，先得知道计算机眼里的声音长什么样。声音的本质是空气的振动。麦克风把振动变成电压起伏，每隔一小段时间记录一次电压值（这叫**采样**，常见频率 44.1 kHz，即每秒记录 44100 次），就得到一串数字——**波形**（waveform）。计算机存储、传输、合成音频，处理的都是这串数字。

但直接处理波形很难"理解"：它只是电压曲线，看不出"谁在说话、说了什么"。于是有了第二种视角——**频谱**：把一小段时间的波形分解成"由哪些频率成分组成"（这份工作由傅里叶变换完成）。人耳的听觉恰好是按频率组织的，所以"频谱"比"波形"更接近"我们听到的内容"。音乐播放器上的跳动柱状图、语音识别与语音合成里的特征，用的都是频谱视角。

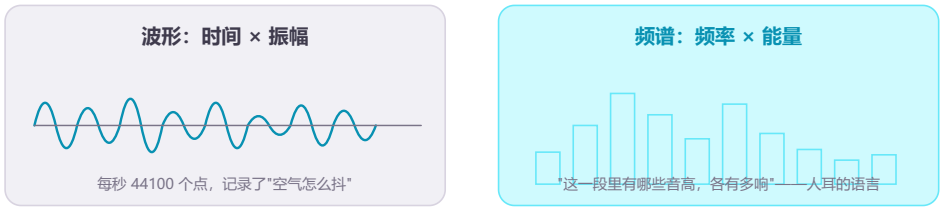


图 1-12　声音的两副面孔。波形是"机器记录的形态"，频谱是"接近听觉的形态"。AI 声音生成的技术管线，正好走在两者之间。

◆ TTS 的三段管线：念得对、念得像、听得真

语音合成（TTS）走的是"文本 → 声音"的路线，现代系统大致分三步：

1. **文本前端：把字读对。**处理多音字（"明月"的"月"没问题，"长安"要读 cháng 还是 zhǎng？）、数字读法（"2 分钟"读"两分钟"）、断句与轻重音——这一步决定了"念得对不对"。
2. **声学模型：把字变成"声音的参数"。**决定用什么音色、什么语调、哪里停顿、哪里扬起来。这一代模型是神经网络，学的是"文本与语音的对应关系"。
3. **声码器：把参数变回波形。**把声学参数"还原"成能播放的波形数据——决定音质是否自然、有没有金属味。

近年的关键进展是**零样本语音克隆**：只要给模型一小段某人说话的录音（几十秒量级），它就能提取"音色特征"并模仿其声音朗读任意文本。技术上它是"把音色作为条件"的 TTS——听感上，它意味着每个普通人都能"拥有"任何声音。

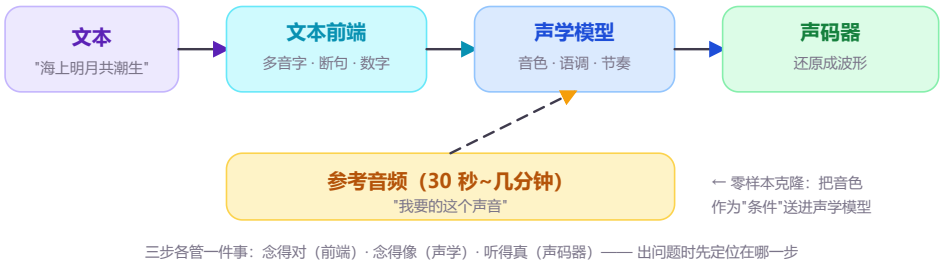


图 1-13　TTS 的三段管线与语音克隆。记住这三段，故障就有的放矢：读错字是前端问题，语气不对是声学问题，音质发闷是声码器问题。克隆功能让"音色"变成了一种可输入的参数。

> ⚠️ **声音克隆的红线**
>
> 《中华人民共和国民法典》明确规定，对自然人声音的保护**参照适用肖像权保护的有关规定**。这意味着：未经许可克隆他人声音（尤其是知名配音演员、歌手、公众人物），可能构成侵权；用克隆声音生成虚假内容传播，还可能触犯更严重的法律。**课堂练习只用：① 自己的声音；② 已获授权的声音；③ 平台提供的公共音色库。**教材在第 6 章（身体影像）与附录 E（合规清单）会继续展开这条主线。

回到脚本上——这是小江在主线任务里的下一站。她的解说词要交给 TTS 变成旁白，而"念得对、念得像、听得真"这三关，恰好对应她要做三次检查：字读对了吗、语气对不对、音质能不能用。

图 1-14　配音的情绪参照。好的旁白不是「读字」，而是「有人在说话」。在脚本会上，小组用这样一张参考图统一「解说者」的情绪基调：安静、辽阔、略有迟疑——它随后被写进配音的情绪提示词。 AI 生成

从脚本到成音：一条完整的制作流水线

把本节前面的零件装成一条线，就是"给短片做旁白"的完整流程。这条线你会在第 3 章（与画面合流）和第 8 章（交给智能体自动化）各重逢一次，值得现在就打牢：

图 1-15　脚本到成音的流水线。"分段生成"是这条线的枢纽：它既是质量控制手段（问题定位到段），也是成本控制手段（重做一段而非全片）。

怎么评价一段 AI 配音：三个维度

听一段配音说"像不像真人"太笼统。实践中用三个可操作的维度来评估（游戏与语音行业用类似的框架，教材做了简化）：

表 1-9　AI 配音的三维评价法

维度	听着好是什么样	常见失分	改善手段
自然度 naturalness	韵律自然，停顿与换气像真人；听久不累	机械感、金属音、字字等距	换更强的声码器工具；文本改口语化；语速 0.9×
表现力 expressiveness	情绪对味：抒情段落柔和、叙述段落沉稳；有轻重缓急	全程一个调；该停不停	选带情感控制的音色；用标点分级控停顿；分段生成
相似度 similarity	（仅克隆场景）像不像目标音色，且稳定不"漂"	长文本中途音色变化；情感一激动就不像了	参考音频选干净、时长充足的；一次生成别太长；分段后统一音色参数

注意三维之间常常**互相拉扯**：追求夸张情绪可能牺牲自然度，追求高度相似可能牺牲表现力。此时回到"用途"做决定——旁白追求自然和稳定，广告配音追求表现力，角色配音追求相似度。你的期末作品属于哪一类，评估标准就向哪一类倾斜。

> 🩺 **失败诊所 F1-3**　"一口气读完"的长句事故
>
> **症状：**小江把一段 120 字的解说词整段粘贴进 TTS，生成的音频像"没有标点的贯口"——中间不停顿，句尾不落音，听着喘不过气。
>
> **成因：**文本前端对长句的断句是"尽力而为"：当一句话里逗号少、修饰成分层层嵌套时，模型不知道在哪里换气。另外，整段一次性合成时，模型缺少"段落感"——它不知道哪里是结尾、哪里该放缓。
>
> **处方：**①**为配音而写文本**：把长句拆成短句，书面语的顿号、"的"字长定语改成口语节奏（"江畔的、曾经照过无数先人的那轮月亮"→"江畔的月亮，照过多少前人"）；②用标点分级控制停顿：逗号=短停，破折号=中停，句号=长停，必要时换行=换气；③分段生成再拼接，比整段生成更容易控制；④生成后用 0.9× 语速回听一遍，凡是"你自己念着别扭"的地方，改文本而不是怪工具。配音质量的上限，往往取决于文本的"可朗读性"，而不是音色。

◆ 拿到音频之后：几个工程参数

AI 生成的音频最终要进剪辑软件、要与其他素材对齐，这时几个基础参数必须心里有数（剪辑软件里都叫这些名字）：

表 1-10　音频工程参数速查

参数	含义	实践建议
采样率 sample rate	每秒记录多少个点，单位 kHz	语音 ≥22.05 kHz 即可；音乐与视频工程统一 44.1 kHz 或 48 kHz，**全片统一一个值**，不要混用
位深 bit depth	每个采样点用多少位存振幅	生成与剪辑用 16 bit 足够；需要多次处理/混音时用 24 bit 留余量
文件格式	WAV 无损体大；MP3/AAC 有损体小	**剪辑与存档用 WAV**，最终交付再导出 MP3/AAC；反复"MP3→WAV→MP3"会叠加损失
声道	单声道 / 立体声	旁白用单声道即可（还能减小体积）；配乐用立体声更有空间感
响度 loudness	整体音量感受（常用 LUFS 衡量）	旁白与配乐要"配平"——旁白听得清、配乐退在其后；发布平台一般有响度建议值，导出前统一调整

1.8 音乐与音效生成：第四条产线

声音这条线上还剩最后一族技术——**音乐与音效生成。**它和 TTS 都属于"音频生成"，但训练的数据、要解决的问题、以及工具生态完全不同：TTS 学的是"人怎么说话"，音乐模型学的是"旋律与和声怎么组织"。这也是我们把它单独列一节的直接原因：**同属"生成音频"，但它们是两个技术族。**

◆ 两条技术路线，与"唱出来"的难点

音乐生成有两条主要路线。一条是**符号音乐生成**：模型先"写谱"（输出 MIDI 音符序列），再由音色库演奏出来——好比先写乐谱、再找乐队。另一条是**直接生成音频**：在音频压缩后的离散编码上建模，像语言模型生成文字一样生成声音——它更接近"直接唱给你听"。

后一条路线解决了传统难题里最棘手的一个：**歌词与旋律的对齐**。人声歌唱要求每个字落在正确的音符上，而"先写谱再演奏"的做法很难精确控制这一点；直接生成音频的模型则可以把歌词和曲风一起作为条件，一次生成完整的"演唱"。这也是近两年"AI 唱中文歌"质量跃升的技术原因。

音效生成（风声、水声、脚步、开关门……）与音乐生成同源——都是"以描述为条件的音频生成"，只是训练数据换成了海量的环境声与拟音素材。

◆ 配乐生成实操：把"感觉"翻译成参数

以"给短片做配乐"为例，一条够用的需求描述通常包含五要素：

表 1-11　配乐生成的需求五要素（《春江花月夜》示例）

要素	说明	示例写法
用途	纯音乐还是有唱词？计划在哪一段使用	纯器乐，用于开场 60 秒旁白底衬
情绪走向	不要只给一个形容词，给出"从…到…"的变化	由静谧微凉，到开阔明亮；结尾回到安静
乐器与编制	主奏乐器、是否用打击乐；规模感	箫或古琴主奏，少量低音弦乐铺底，不用鼓
时长与结构	精确到秒；有无渐强/收束要求	60 秒整，末尾 3 秒自然收束（渐弱不截断）
禁忌项	明确不要什么（比"要什么"更能救回结果）	不要人声、不要电子鼓点、不要突兀转调

音乐生成最常被低估的一点：**"时长精确到秒 + 自然收束"是它相对传统曲库的最大优势**。传统做法要"剪辑到秒"，常常破坏音乐的完整结构；而生成工具可以"因片制宜"地写音乐。反过来，它最不擅长的也是传统曲库的强项：有主题记忆点的旋律——"让人哼得出来"的主题，目前的生成工具很难稳定提供，这也是"人工修改"在配乐环节的用武之地。

◆ 音效与拟音：让画面"落地"的隐形功臣

如果说配乐决定"情绪"，音效（sound effect）决定的就是"真实感"：江面的水声、衣袖摩擦的窸窣、远处的钟鸣——这些东西观众注意不到，但一旦缺失，画面立刻"塑料感"。电影工业里有专门的工种叫**拟音**（foley）：拟音师在棚里用真实道具"演"出画面的声音（哗啦的水声可能来自揉搓的塑料袋）。

AI 音效生成把"找素材"变成了"写描述"：输入"夜晚的江水轻拍石岸，远处一声孤钟"，直接得到可用的音效。它对短片创作的最大价值有三：① **描述即所得**——传统音效库检索靠关键词碰运气，现在是"你想要什么就说什么"；② **时长可控**——和配乐一样可以按场景精确到秒；③ **氛围一致**——同一段落的多个音效可以在同一"描述风格"下生成，听感统一。局限同样明显：要求"真实录音级"的复杂音效（如特定型号发动机的轰鸣、某样乐器的特殊技法）时，生成结果常常"像但不真"，这种场合仍需真实素材库与拟音。第 3 章的剪辑环节会再回来讲音画配合，届时你会用上这里的工具。

> **⚠️ 音乐类工具的条款差异最大**
>
> 与 TTS 不同，音乐与音效类工具在**生成内容的权属与商用许可**上差异极大：有的平台把生成物版权按比例与创作者分成，有的仅授予个人非商用许可，还有的在国际版权诉讼影响下频繁调整条款（下载权限、商用范围都会变）。**课堂作业请只使用"明确允许非商用创作"的档位**；若作品要参加比赛、发布到公开平台或用于商业场合，请回到该工具的官方条款逐条确认——这一步在第 8 章的"发布前验收清单"里还会出现。

> **📌 声音这条线：四族技术到此收束**
>
> 回头看本章的四族技术：**大语言模型**（写文字）、**语音识别 STT**（声音→文字）、**语音合成 TTS**（文字→声音）、**音乐与音效生成**（以描述生成音频）。它们共享"语言与声音"这条介质线，也共享同一套工作方法：把需求说清楚（提示词与参数）、把结果评出层次（三维评价、词错率、五要素）、把人工放在关键闸门（核对、试听、条款确认）。下一节把这些方法收进一张工具地图，并给出这一章最重要的合规底线。

1.9 工具地图与合规框架

原理讲完，落到"用什么工具"。本节的两张表 + 一张图，是你第 1 周环境准备的行动清单。工具迭代极快，本书的原则是：只记"能力类别"，不迷信具体型号——附录 B 会持续维护最新的工具清单。

图 1-16　文本与声音的能力地图。不背工具名，先认能力格：需要哪一种能力，再去附录 B 找当期可用的具体工具。

表 1-12　声音类工具速查（据编写本书时的公开资料整理；额度与条款以官方页面为准）

工具	能力	门槛与额度（本时点口径）	课堂适用场景
剪映（字节）	AI 朗读/配音、数字人；与剪辑无缝衔接	基础配音功能免费、门槛最低；与画面对齐方便	**首选实践入口**：配音 + 剪辑一条龙，机房环境友好
讯飞智作（科大讯飞）	配音 + 数字人；中文自然度与方言适配口碑好	免费版有声线/时长限制；商用需开通会员确认授权	需要更自然的中文语气与方言时
火山引擎豆包语音（字节）	语音合成、声音复刻（克隆）、音色设计	语音合成大模型**仅对已认证企业用户开放**，个人课堂较难直接使用；克隆需提供合法语音样本与授权	了解"企业级"音色工程的样板；个人作业不建议依赖
CosyVoice（阿里通义，开源）	零样本克隆（约 3–10 秒参考音频即可复刻音色与韵律）、多语种与方言、情感指令控制	**开源可免费本地部署**；适合有显卡、想深入技术的同学	附录 F 的选学路径；理解"克隆为什么这么像"的最佳教材
MiniMax 语音/音乐（海螺）	语音合成、定制音色、音乐生成	开放平台按量计费；网页端有体验入口	需要"歌唱/哼唱"类生成时
天工 SkyMusic（昆仑万维）	音乐生成：一句话或歌词成曲，支持参考音乐与方言歌曲	网页/App 入口；额度以官方为准	短片配乐实验；注意条款对商用与权属的表述
Suno / Udio（国际对照）	整首歌生成，质量口碑领先	需海外网络；免费档多限个人非商用；两家均在 2026 年前后经历**授权模式转型**（与唱片公司和解、下载功能调整），条款变动频繁	作为"能力强但条款复杂"的案例讨论；不用于课程交付

看这张表要抓的重点不是"哪个最好"，而是三种**门槛类型**：**身份门槛**（要不要企业认证）、**费用门槛**（免费额度够不够练习）、**条款门槛**（生成物的权属与商用许可）。课程实践的策略很朴素：用"免费额度 + 公共音色 + 自己的声音"完成全部作业；把商业级工具与开源方案留作拓展。

◆ 模型家族：认识"选手"，而非背"比分"

市面上的大模型看似眼花缭乱，其实是几个"家族"在竞争。下表按"厂商 → 家族"组织——版本号几个月一换，但家族的技术路线与擅长领域相对稳定。表中型号为编写本书时（截至 2026 年 10 月）检索到的版本，仅供建立"格局感"，使用前请以各官方文档为准：

表 1-13　国内主流大模型格局（截至 2026 年 10 月）

厂商	家族与代表版本	本时点特点	对创作的意义
深度求索 DeepSeek	V4 系列（Pro / Flash）	代码与数学推理见长、成本低；超长上下文	脚本结构分析、工作流规划、技术说明撰写
阿里 · 千问 Qwen	Qwen3.5 系列（开放权重）	原生多模态（可读图、读视频）；语言覆盖广；开放权重便于本地部署	多模态素材分析；"本地跑模型"（附录 F）首选之一
字节 · 豆包	Seed 系列（Pro / Lite / Mini）	与字节系创作工具（图像/视频/剪辑）生态打通；多档位匹配不同难度	一条龙创作流：文案 → 图像 → 视频 → 剪辑
月之暗面 Kimi	K3 系列	超长上下文与长文档处理见长	整本资料、长剧本的阅读与校对
智谱 GLM	GLM-5 系列（开放权重）	开源可本地部署；适配国产芯片生态	体验"技术自主"路线；研究性学习
百度 · 文心	文心 5.0 系列	原生全模态统一建模	中文语境写作与检索增强问答
腾讯 · 混元	混元系列（元宝入口）	与微信生态绑定；三维与视频能力见长（第 4 章详述）	社交场景分发；三维资产生成
MiniMax（海螺）	M 系列 + 语音/音乐专用模型	语音与音乐生成是强项（见 1.7、1.8 节）	配音、配乐的进阶选择

表 1-14　国际旗舰模型（对照参考；中国大陆多有访问限制）

家族	特点	课程中的定位
GPT 系列	综合能力前沿，多模态与工具生态成熟	"能力上限"的参照物；技术说明中可作对照
Claude 系列	长文本、代码与写作细腻度口碑突出	写作质量的参考系
Gemini 系列	多模态融合深，长上下文与视频理解见长	多模态任务的参考系
Llama 等开源	社区生态庞大，可本地部署	研究性学习；英文任务为主

> ⚠️ **关于"能不能用"的如实说明**
>
> 本书实践环节以**国内可直接访问的网页版工具**为主线（这也是课程机房环境的要求）；国际模型在部分任务上仍是能力参照，但存在访问门槛，使用需遵守所在地与平台的相关规定。**课程作业不要求、也不禁止使用特定工具**——要求的是"用你能合法、稳定访问的工具，把方法做出来"。

一个可以写进技术说明的背景数字：据公开报道，2026 年中国大模型的周调用量已连续数十周位居全球最活跃梯队，应用市场的规模与迭代速度都处在世界前列——你在这门课里练的，正是这个市场里最紧缺的技能：会用、会评估、会负责。

◆ 开源意味着什么：不只是"免费"

表中"开源派"值得单独说两句，因为它是中国团队对世界 AI 生态最实在的贡献之一，也是通识课学生最应该理解的"技术自主"概念：

- **可下载**：模型权重公开，任何人可以在自己的机器上运行——数据不出本地，适合对隐私敏感的场景。
- **可改造**：可以针对特定任务（比如"专门给诗词短片写脚本"）做微调，让模型更像"你的助手"。
- **可审计**：代码与权重公开意味着安全研究可以独立进行——这对整个生态的信任至关重要。
- **成本可控**：本地运行不按次付费；配一块显卡就能跑 7B–30B 级别的模型（具体门槛见附录 F）。

对这门课而言，"开源"还有一个更直接的意义：**它让"技术自立自强"从一句口号变成你桌面上可以运行的东西。**第 8 章讲智能体时，你会看到开源模型如何作为"本地大脑"与云端工具协作。

◆ 合规框架：从"三件套"到"全套装备"

中国对生成式人工智能的治理已经形成清晰的法规框架，并且还在密集更新。作为创作者，至少要认识下面这张表里的规则（细节与最新动态见附录 E）：

表 1-15　面向创作者的合规框架（截至 2026 年 10 月，简版）

法规/文件	管什么	与你创作的关系
《网络安全法》（2026 年 1 月 1 日修订施行）	首次将人工智能写入法律：支持发展，同时要求完善伦理规范、加强风险管理	理解了它，就理解了整个治理框架的基调：**发展与安全并重**
《生成式人工智能服务管理暂行办法》（2023 年 8 月 15 日施行）	生成式 AI 服务的提供与使用	你使用的工具应依法备案；使用时要遵守平台规则，不得生成违法有害内容
《互联网信息服务深度合成管理规定》（2023 年 1 月 10 日施行）	深度合成技术：换脸、**合成人声/仿声**、智能写作、虚拟场景等	配音、数字人、AI 写作都落入"显著标识"范围；不得用于伪造他人
《人工智能生成合成内容标识办法》+ 强制性国标 GB 45438—2025（2025 年 9 月 1 日施行）	AI 生成合成内容的**标识**：显式 + 隐式双轨；国标给出量化要求（如图片显式标识文字高度不低于画面最短边的 5%，文字须含"AI/人工智能"与"生成/合成"要素）	**创作者最直接的操作义务**：发 AI 内容要"亮明身份"，且标识有最低技术要求
《学位法》（2025 年施行）	首次将"利用人工智能代写学位论文"纳入学术不端范畴	课程红线：AI 可以辅助，不能代写；这门课的"技术说明"正是在训练"辅助"的正确姿势
《人工智能拟人化互动服务管理暂行办法》（2026 年 7 月 15 日施行）	面向"模拟人类人格特征与沟通风格的情感互动服务"（AI 聊天、虚拟伴侣等）	你的作品若包含"AI 角色与观众互动"，就落入本办法的监管范围
最高法《关于依法审理涉人工智能纠纷案件的意见》（法发〔2026〕10 号）	首部由最高审判机构发布的涉 AI 裁判规则文件，24 条	看"司法怎么看"：声音权益、生成内容标识、数据使用边界都有了裁判口径

记忆法：把这张表压缩成一句话——用合法的工具、标清楚身份、不碰他人权利、不代替自己思考。四句话分别对应上面的法规群：备案准入、标识义务、人格权与个人信息、学术诚信。第 2 章的"素材来源"、第 6 章的"身体影像"、第 8 章的"透明协作"会沿着同一根主线继续展开。

案例 1-B § 24 个共振峰：两起判决把"声音"关进了笼子

背景：两起被媒体称为"标杆"的判决。其一（2024 年，北京互联网法院）：配音师殷某发现自己的声音被 AI 化处理后，出现在某文本转语音产品里供人购买使用——源头是此前录制的录音制品被转卖训练；法院认定"经 AI 合成的声音若能让公众关联到特定自然人，即受法律保护"，并明确"**对录音制品的授权并不意味着对声音 AI 化的授权**"，判赔 25 万元。其二（2026 年 9 月，上海一中院二审终审）：配音演员王女士发现某推广活动使用了疑似其声音的 AI 合成音频；司法鉴定显示，28 个共振峰声学指标中 24 个偏离度小于 10%、高度相似部分达 90%，法院判侵害声音权益成立，判赔 5 万元。

关键细节：第二案中，被告公司辩称"音频确系 AI 生成，但训练素材来源因前员工离职无法确认"——法院明确：**作为掌握原始训练数据的控制方，未能证明素材来源合法，应承担举证不能的不利后果。**

复盘：这两起判决给创作者的三个提醒：①"大家都能听出来是你"就够了——法律保护的标准是**可识别性**，不是"逐样本完全相同"；②合法拿到录音，不等于有权拿它训练 AI 或克隆声音，授权要写明用途；③"素材来源说不清"本身就是最危险的状态——这也是本课程从第 1 周就要求建立"素材来源档案"的原因。法官的一句话可直接记入手册："技术中立不等于使用中立，AI 模仿人声并非原创。"

留给你的一句话：技术门槛越低，法律意识的"地板"越要抬高。工具箱里永远有一个不花钱、零风险的选项：用你自己的声音。

案例 1-E § 文旅短视频的"三稿工作流"

背景：某地方文旅账号要发一条 60 秒的短视频，介绍一处宋代古桥。要求"文案有文化味但不掉书袋，配音要亲切"。

做法：运营者建立了"三稿工作流"：**一稿求全**——给模型尽可能多的背景材料（历史沿革、建筑特征、民间传说），要求它先列要点清单，不写成文；**二稿求准**——对要点清单逐条核查（年代、数据、传说是否与方志一致），删掉无法核实的两条，让模型按确认后的要点成文，并指定"每句 ≤18 字、不用成语堆砌"；**三稿求味**——把定稿的文案作为示例，让模型"保持这个味道"另写开头句和结尾句各五版，人工挑一版。

结果：从选题到成片文案约 90 分钟，其中人工时间（核查与挑选）约占一半。视频发布后，评论区夸得最多的是结尾那句——那是"三稿求味"人工挑选出来的。

复盘：①"先列要点、再成文"结构上把"查证"和"创作"分开了，避免了"流畅的句子里埋着错误"；②"人工时间占一半"不是效率低，而是**把人的时间集中在了判断力上**；③"让模型给候选、人来挑"比"让模型一步到位"更可靠——这也是本书后面所有媒介的通用工作方式。

🩺 **案例 1-D** § 给视障读者做一本"AI 有声书"

背景：某高校志愿团队想为视障读者制作有声读物，选材是地方志里的民间故事。传统做法需要请人录制，成本高、周期长。

做法：团队用 TTS 生成初版音频，但很快发现"直接念"远远不够：①方志里的文言引文，机器断句生硬；②人名、地名里的多音字错误率高；③故事中对话与叙述的语调没有区分。于是他们把流程改造为"AI 起草 + 人工精修"：先让大语言模型把文言引文改写成口语，再由人工逐句标注多音字与停顿，最后分段生成、人工拼接。

结果：制作周期从"数周录音"缩短到"数天"，人工精修的部分（约 20%的句子）恰恰是最影响听感的 20%。作品在读书日活动中发布，并按要求标注了"AI 语音合成"。

复盘：三个经验值得记住：①**AI 把成本从"录制"转移到"精修"**，人力不是被省掉了，而是被重新分配到了更需要判断力的环节；②无障碍场景对准确性的要求极高（读错一个人名，对读者是伤害），因此"多音字词表 + 人工校对"不可省略；③这类公益性用法是生成式 AI 最值得被鼓励的方向之一——技术向善不是口号，是具体的选题。

◆ 数据与隐私：你粘贴进去的内容去哪儿了

一个必须直面的问题：你在对话框里粘贴的每一段文字、上传的每一张图，去了哪里？不同的服务条款给出不同的答案，而你需要形成两个习惯：

- **先看条款再粘贴。**多数在线服务默认可能将对话用于"服务改进"（含义常常包括训练）；是否可关闭、是否可用于训练、保留多久，各家不同。凡是不能公开的素材（他人隐私内容、未发表作品、含个人信息的数据），不要粘贴进未确认条款的在线服务。
- **敏感素材走"本地路线"。**这正是1.9 节"开源模型"的现实价值：本地部署的模型推理在你自己机器上完成，数据不出设备。附录 F 给出了一个可复现的本地演示路径（本课程不要求，但值得知道它存在）。

还有一层"隐私"容易被忽略——**你自己作品的隐私**：期末作品在正式提交前，把关键创意（完整分镜、成片）上传到大流量平台求"点赞"，可能带来作品被抢先发布的尴尬。课程鼓励同学间互评，但建议在班级范围内进行；对外发布一律等作品定稿、标识齐全之后。

◆ 平台生态：把单点工具连成"创作流"

到你实际操作时就会发现：效率的瓶颈很少是"单个工具不够强"，而是"工具之间的搬运"。写完的文案要复制到配音工具，生成的音频要下载后导入剪辑，剪好的片段要重新上传——每一步手动搬运都在消耗注意力。**平台生态**正是在解决这件事：把对话、写作、图像、视频、配音、剪辑放进同一个账号体系，"上一环节的产物 = 下一环节的素材"。

国内主流平台几乎都走了这条路：聊天产品内置图像/文档/语音能力；创作平台（如以图像视频见长的国产工具）把生成与编辑、社区与素材库打通；剪辑工具把"文本成片、智能字幕、配音"集成进时间线。对课程实践来说，选平台的一条实用标准就是：能不能覆盖你需要的"至少两类媒体"——期末作品要求融合两类以上媒体，能在同一平台内完成的组合，出错率和搬运成本都最低。当然，也别被生态绑死：关键资产（提示词、参数、素材原文件）始终要保存在你自己的创作档案里，这既是第 1 章就强调的习惯，也是未来换平台时的底气。

◆ 选型三问：怎么在几十个同类工具里做决定

工具永远在换，但"选型"的方法稳定。面对一类能力（比如"文本转语音"），问三个问题就能筛出可用的：

1. **中文支持怎么样？**看它对中文语境的处理：多音字、断句、语气词；中文样本的听感是否自然。国际工具未必弱，但要实际试听中文效果再决定。
2. **免费额度与限速是多少？**课堂练习需要"能反复试错"的额度：每天几次、每次多长、有无排队。额度小的工具适合"精修"，额度大的适合"批量起草"。
3. **商用条款与授权怎么写？**这是最容易被忽略、后果最严重的一条：生成内容的版权归谁？能否商用？是否需要署名？训练数据是否包含他人作品？——**课程作业也要养成查条款的习惯**，因为期末作品要公开展示（属于"对外发布"）。

把三问的答案记进你的创作档案。附录 B 的工具速查表按这套格式维护：能力 → 工具 → 中文支持 → 免费额度 → 条款要点。工具会变，**方法不会变**。

1.10 关键概念表

本章出现的技术词汇，在此集中"对账"。建议把这张表当作本章的复习清单：能用自己的话解释每一行，本章的原理目标就达成了。

表 1-16　第 1 章关键概念（中英对照）

术语	英文	一句话解释
模态	modality	信息的形态：文本、图像、音频、视频、三维等
多模态	multimodal	把不同模态编码进同一向量空间，统一处理
词元	token	模型处理文本的最小切块单位，介于字与词之间
分词器	tokenizer	决定文本如何切词元的规则/程序
嵌入	embedding	把词元/图像块映射成向量，语义相近则距离相近
余弦相似度	cosine similarity	用向量夹角衡量"含义有多接近"，第 8 章检索的基础
大语言模型	LLM	以"预测下一个词元"为核心任务训练的大规模神经网络
Transformer	Transformer	2017 年提出、以大模型为主流的网络架构
注意力	attention	处理每个词时给上下文分配权重的机制，本质是加权平均
预训练	pretraining	在巨量文本上学习语言规律的第一阶段
指令微调	instruction tuning / SFT	用"指令→回答"数据教会模型"听话"
人类反馈对齐	RLHF / alignment	用人类偏好教会模型"什么叫更好的回答"
上下文窗口	context window	模型一次能读入的词元上限；模型默认无长期记忆
推理模型	reasoning model	先写长草稿再作答的模型，适合多步复杂任务
提示工程	prompt engineering	设计输入以稳定获得期望输出的方法
零样本/少样本	zero/few-shot	不给示例 / 给少量示例的提示模式
温度	temperature	控制取样随机性的参数；低=稳，高=多
核取样	top-p	只在累计概率前 p 的候选里取样的过滤参数
幻觉	hallucination	流畅自信但不符合事实或凭空编造的生成内容
语音识别	STT / ASR	把语音转成文字的技术；与 TTS 互为逆过程（判别 vs 生成）
词错率	WER	识别准确度的行业指标：错词数 ÷ 总词数
热词表	hotwords	提前告知模型的重要词汇（人名/术语），识别前最划算的一步
说话人分离	speaker diarization	区分"谁在什么时候说"
语音合成	TTS (text-to-speech)	把文本转成语音的技术
波形 / 频谱	waveform / spectrum	声音的时域记录 / 频率构成表示
声码器	vocoder	把声学特征还原成波形的模块
语音克隆	voice cloning	用少量参考音频模仿音色朗读任意文本
深度合成	deep synthesis	法规语境下对换脸、语音合成等技术的统称
显式/隐式标识	explicit / implicit marking	AI 生成内容的可见标识 / 元数据标识（法规要求）

1.11 动手实践：从环境准备到第一段配音

🛠 实践一：环境与账号准备（30 分钟 • 课后完成）

基础

按附录 B 的清单，注册并登录至少：1 个通用对话助手、1 个语音合成/TTS 工具、1 个音乐或音效生成工具，以及 1 个网盘/笔记工具（用于存放作品与生成记录）。

提交物：一张工具清单截图 + 每类工具记一条"我用它做什么"。

进阶

建立你的"创作档案"文件夹结构：作品 / 参数记录 / 对照实验 / 素材来源。为每个素材建立来源档案（来源 URL、授权方式、下载日期）——从第 1 周就用起来的习惯，到期末会是你的救命稻草。

🛠 实践二：提示工程初训——同一任务的两种写法（40 分钟 • 课堂）

1. **写裸提示词。**登录对话助手，输入："写一段《春江花月夜》的解说词。"复制回复，编号为"版本 A"。
2. **用四要素重写。**把本节的提示词卡复制进去（把占位符换成你的具体设定），得到"版本 B"。
3. **对照评估。**从四个维度打分（各 1–5 分）：画面感、可朗读性（是否拗口）、字数控制、是否可用（能不能直接进分镜表）。填入下表并保存——这就是你的第一份"参数对照记录"。

表 1-17　对照实验记录表（模板，复制到你的创作档案）

版本	画面感	可朗读性	字数控制	可用性	备注（差在哪）
A（裸提示词）	–	–	–	–	填写你的评分与理由
B（四要素）	–	–	–	–	填写你的评分与理由

🛠 实践三：参数对照实验——温度（30 分钟 • 课堂）

进阶

用同一个提示词（自选任务，如"给分镜 1 起 5 个标题"），分别以**低温度（≈0.2）**和**高温度（≈1.0，若界面可调）**各生成 3 次，共 6 组结果。观察并记录：

- 低温度的三次结果之间差异大吗？高温度呢？
- 哪种温度更适合"找灵感"，哪种更适合"要稳定"？

提交物：6 组结果 + 一段 100 字以内的结论（写进实践记录）。

挑战

让模型用一句话解释它自己在高/低温度下的差别（提示词示例："用一句话解释：为什么你刚才在'温度高'时答案更发散？"），把你认为解释得好的那句话记下来——这也是"用 AI 学 AI"的开始。

🛠 实践四：主线推进——解说脚本、配音与字幕（70 分钟 • 课堂 + 课后）

1. **定稿脚本。**用实践二的版本 B 继续迭代，产出《春江花月夜》前 60 秒的解说词与 4–8 个镜头表（镜头号、画面描述、旁白、建议时长）。
2. **核对事实。**凡引用诗句原文，回原诗逐字核对；凡涉及历史细节的描述，标注"已核实 / 待核实"。
3. **生成配音。**把解说词粘贴进 TTS 工具：选择音色（公共音色库或你自己的声音）→ 处理多音字（"斜"字在古诗里的读音、"扁舟"的"扁"读 piān）→ 调整语速与停顿 → 导出音频。
4. **质检。**戴上耳机逐句听：有没有读错的字、不自然的停顿、过重的机械感？对照音频与文本，把问题分成"文本前端问题"（读错）与"声学问题"（语调），再针对性调整。
5. **做字幕。**把配音音频拖进语音识别工具：先建**热词表**（"春江花月夜""滟滪""汀上白沙"等），再识别；人工逐句核对同音字与专有名词；导出带时间戳的字幕文件，导入剪辑软件生成字幕轨。
6. **归档。**把音频、脚本、字幕文件、生成参数（工具名、音色、语速、日期）存入创作档案，并给音频加 AI 生成标识说明。

常见问题 A • 多音字读错

症状："江水流春去欲尽"的"尽"读成 jìn（应为 jǐn）。

处方：把多音字替换成同音字或拼音（如写作"去欲紧"生成后再改回），或使用支持"拼音标注"的工具。

常见问题 B • 语气平淡

症状：整段一个语调，像导航播报。

处方：在文本里加"停顿标记"（换行、逗号、破折号分级）；把长句拆短；换一个"新闻/纪录片"类的音色风格；语速降到 0.9×。

🛠 实践五：配音参数对照实验（40 分钟 • 课堂）

进阶

固定同一段文本（用实践四的解说词第一段），只改变**一个**变量，各生成一版并对比：

表 1-18　配音 A/B 对照实验设计

对照组	变量	A 版	B 版	观察点
实验一	语速	1.0×	0.85×	抒情段落的"呼吸感"哪个更好
实验二	音色类型	年轻女声	沉稳男声	哪一类更贴《春江花月夜》的意境
实验三	文本标点	原文（长句）	拆成短句 + 换行	停顿是否更自然

提交物：三组 A/B 音频 + 每组的结论一句话（形如"实验三表明：为配音改写标点比调语速对自然度的提升更大"）。

挑战

用音乐生成工具为主线的 60 秒段落生成一版配乐：情绪"由静到明"，时长精确到 60 秒且结尾自然收束。再生成一版"错误示范"（不写时长、不给情绪），对比两者的可用性——感受"约束"对生成质量的影响。

1.12 课程思政：文化传播与声音权益

本章的思政主题有两个：**科技向善与负责任的使用**，以及**声音权益与文化传播。**

中国贡献：这轮技术变革里的"中国方案"

生成式 AI 不是某一家公司的独角戏。中国团队在这一轮竞赛中贡献了几样世界级的公共品，而且都可以核查，不是口号：

表 1-19　「中国方案」的三样公共品（据公开资料整理，截至本书编写时）

公共品	具体是什么	为什么重要
高效架构	以混合专家（MoE）等路线，用远低于同级的算力成本达到第一梯队能力	把「用得起」变成现实——中小团队与个人也能做出像样的作品，这是创作普惠的前提
开源生态	多家中国团队开放模型权重与工具链（有的以宽松许可发布，允许商用与二次开发）	全球开发者可以自由下载、本地部署、按需改造——「可利用」比「能调用」更根本
应用广度	把 AI 接进制造、医疗、教育、文旅等真实场景；生成式 AI 服务备案数量持续增长，中国大模型的周调用量长期居世界最活跃梯队	技术只有进入真实场景才算落地——你在这门课里练的，正是这个市场最缺的技能

从"跟跑"到"并跑"再到局部"领跑"，最有说服力的不是宣传，而是你手里的工具：这门课后面每一章的实践，用的都是这场竞赛的成果。理解"中国方案"，最好的方式是把它用出作品来——这也是本课程把大量实践放在国产工具上的原因。

> 🇨🇳 **课程思政 · 用声音讲好中国故事**
>
> 声音是文化的载体：方言里的乡愁、戏曲里的韵白、诗词吟诵里的平仄。生成式语音技术一方面让"每个人的声音都能被复制"，另一方面也带来了"文化声音被滥用"的风险——试想，如果用克隆音色让某位戏曲名家"演唱"一段他从未唱过的戏词，流传出去，受到损害的是谁？**写脚本时尊重文化语境的准确性**（不把唐代的意象写成宋代的、不混淆方言的文化归属）；**用配音时守住声音权益的边界**（授权、标识、不冒充）；**做传播时坚持诚信**（不把 AI 合成说成真实录音）。三条准则，落到实处就是每一次导出文件时的自查。

声音权益：从"知道有红线"到"三条可执行的习惯"

第 1.7 节讲过克隆的红线，第 1.9 节给过两起判例（「录音制品授权 ≠ 声音 AI 化授权」「举证不能者担责」）。知道红线还不够，落到实处只有三条习惯，请从这一周就开始执行：

1. **只用三种声音来源**：① 你自己的声音；② 平台公共音色库；③ 有书面授权的声音。除此之外的第四种来源——"网上找的""听着像谁就用谁"——一次都不要碰。练习时多花三十秒找公共音色，比事后道歉便宜得多。
2. **导出即标识**：AI 合成或克隆的音频，导出时就带上「AI 生成/合成」的说明，不留到发布前才想起来（第 3 章会给出视频版的完整标识流程）。
3. **保留来源凭证**：授权书、平台条款截图、生成记录一并存入创作档案——真出事时，"我说得清来源"就是最有力的证据；说不清来源，才是判例里最危险的状态。

学术诚信：这门课"辅助"与"代写"的边界在哪

2025 年施行的《中华人民共和国学位法》把"利用人工智能代写学位论文"明确纳入了学术不端范畴——这不是遥远的新闻，它划定了每位同学与 AI 相处的法律边界。落到这门课上，"辅助"与"代写"的边界可以这样判断：

表 1-20　"辅助"与"代写"的判断标准（本课程适用）

判断问题	属于"辅助"（鼓励）	滑向"代写/代做"（禁止）
你能解释每一步为什么这么做吗？	能：参数、取舍、修改理由都在你的掌控中	不能：结果从哪来、为什么这样，说不清
过程记录存在吗？	提示词、参数、对照实验、修改痕迹完整	只有最终成品，过程"不可复现"
关键判断是谁做的？	候选由 AI 给，取舍与验收由人做	全盘接受 AI 输出，不做评估与修正
AI 参与比例如实说明了吗？	提交材料中如实标注所用工具与人工修改	隐瞒 AI 参与，把生成物当"原创"提交

你会发现，这四条判断标准与前面"五步工作法"完全同构——课程的每一项设计（参数记录、对照实验、技术说明、个人贡献声明）本质上都是在帮你证明"我在辅助的这边"。记住一句话：允许 AI 参与你的思考过程，但不允许它替代你的思考过程。

另一面：同一项技术，三个值得做的方向

前面几节讲的多是"用错会怎样"。把镜头转过来——同一套语音技术，有三类用法几乎不可能被质疑，而且正在真实发生：

① 无障碍

把长文与教材转成有声读物，让视障读者"读"到更多；给听障人士做实时字幕。技术在这里的角色是消除门槛。

② 文化保存

为方言、濒危语言、口述史留下可检索的语音档案——很多老人一辈子只被记录过这一次。前提是**知情同意**与**用途限定**。

③ 教育普惠

把优质课程内容转成可听、可跟读的音频，送往师资薄弱的地方。这是"人工智能+"在民生领域的典型落点。

为什么要把这一节和前面的警示放在一起讲？因为技术向善与技术在作恶之间，差的从来不是技术本身，而是使用者的选择。这门课教的每一招，既可以用来让外婆的方言留下来，也可以用来伪造一句她从未说过的话——这两个用法在操作上的差别可能只有几步，在性质上却是两个世界。做选择的能力，比做作品的能力更值得练。

◆ 文化传播的责任：一个"差不多"的错误能走多远

讲一个值得所有创作者警惕的场景。某账号用 AI 生成了一组"古风插画"介绍本地景点，配文里把宋代建筑说成了唐代风格、把地方方言词用错了语境。内容流量很好，评论区却来了几位本地历史爱好者逐条纠正——最伤人的一条评论是："你们连自己拍的地方都没查过资料。"

这个场景里的错误"并不大"：一句含混的年代、一个用错的词。但传播的杠杆是可怕的：AI 让内容生产的成本趋近于零，也让错误的复制与扩散趋近于零成本。过去一个作者查错一处资料，影响几十个读者；现在一条未经核实的生成文案配上精美的 AI 插画，可能被几十个账号二次搬运。技术放大的不是"错误本身"，而是"错误的速度与规模"。

对创作者而言，可操作的结论有三条：**① 涉及历史、地理、民俗的表述，先查权威来源再落笔**（地方志、博物馆、学术资料，而不是"另一个 AI"）；**② 拿不准的细节宁可模糊**（"宋时此地已见记载"优于编造具体年份）；**③ 出错后第一时间更正**——公开更正不是丢脸，是专业素养的一部分。这三条与第 2 章的"素材来源合规"、第 7 章的"文化准确性"是同一套职业标准：**越容易生成，越要负责**。

◆ 技术、职业与劳动：一个诚实的讨论

学到这里，有一个问题值得摊开讲：AI 配音会取代配音演员吗？与其给一个口号式答案，不如看三个事实。第一，**标准化语音确实在被替代**——导航播报、通知提示、批量口播这类"信息传递型"配音，AI 已经做得又快又便宜，这部分工作在减少。第二，**表演性语音的门槛在升高**——需要表演、即兴、与观众建立情感连接的配音（纪录片解说、广播剧、角色配音），观众要的恰恰是"这个人此刻的表达"，AI 的"平均音色"替代不了，但会用 AI 的配音演员在批量试音、小样制作上效率大增。第三，**新的工种在出现**——"AI 声音设计师""提示词撰稿人""生成内容审校"，这些岗位五年前还不存在。

把这三个事实连起来看，得到的不是"取代或不取代"的二选一，而是：技术把行业从"人力密集"推向"判断密集"——被抬高的是"不可替代的那部分人"的价值，被挤压的是"可标准化"的环节。这对每位同学的启示：无论将来做不做媒体，都要问自己一个问题——我的工作里，"可标准化的部分"和"不可替代的部分"各占多少？前者学着交给 AI，后者用心经营。

◆ 劳动与创造：AI 配音背后的人工价值

这一章你亲手做了配音，应该能体会到：AI 生成的每一段"自然"的语音背后，都有人在处理多音字、拆分长句、校对引文、反复试听。技术把"发声"变便宜了，把"发声得好"变成了更稀缺的能力——判断哪里该停顿、哪句话不该那么写、这个音色配不配这首诗。这就是"人的位置"：不做 AI 的对手，也不做 AI 的传声筒，做它的导演。下节课开始，我们进入图像的世界——小江的脚本将第一次"看得见"。

新手常见误区速查（本章复盘清单）

表 1-21　第 1 章易错点与纠正

误区	为什么错	纠正
"提示词越长越好"	关键不是长度，而是"要素完整 + 结构清晰"；堆砌无关信息反而稀释重点	用四要素检查表：角色、任务、约束、输出格式，缺什么补什么
"AI 说的就是对的"	生成是取样而非查证，流畅与正确无关	事实类输出一律核查；不清楚的让它标注"不确定"
"换更好的模型就能解决问题"	多数失败源于需求不清或缺少参考材料，不是模型能力	先按"提示→上下文→参数→模型"的顺序排查，最后才换模型
"温度调到 0 就完全可控了"	低温度只降低随机性，不保证正确性与可复现性	要复现就固定种子；要正确就核查；要稳定就用同一参数记录重跑
"随便找个音色配一下就行"	音色是作品人格的一部分；且音色使用有授权边界	把"选音色"当成一次表演选角；只用公共音色库、自己或已授权的声音
"生成完就完了"	记录的缺失让"技术说明"无从写起，也让经验无法沉淀	每次像样生成都存"提示词 + 参数 + 结果 + 一句复盘"

把这一章压缩成七句话

1. 多模态的本质：一切媒介都先被"切块 + 编码"成向量，住进同一个坐标系——词元与嵌入是全书的"数字地基"。
2. 大语言模型是一台"预测下一个词元"的机器；要预测得准，它必须学会语法、常识与推理——能力由此涌现。
3. 注意力机制是 Transformer 的核心：处理每个词时给上下文分配权重，做一次加权平均；简单机制的成百上千层堆叠，成就了大模型。
4. 训练三阶段："预训练学知识，后训练学行为"——知识有截止日期，听话程度与安全边界来自后训练。
5. 提示词四要素：角色、任务、约束、输出格式（加分项：示例、参考素材）；温度决定随机性，"迭代、一次改一个变量"是基本功。
6. 幻觉的解释与对策：它在取样而非查证——事实类输出必须核查，核查三板斧是换问法、要出处、用检索。
7. 声音生成的三段管线：文本前端（念得对）→ 声学模型（念得像）→ 声码器（听得真）；克隆音色是"把音色变成参数"，但声音权益是法律红线。

自测

❓ 自测 1：为什么说"大语言模型只是在预测下一个词"？这句话是贬低它还是说明它的强大？

答案：这是对它工作方式的准确描述——每一步输出"下一个词元的概率分布"并取样。这不是贬低：要在大规模文本上把"下一个词"预测准，模型必须内化语法、事实、逻辑乃至推理模式，因此"预测"任务训练出了"理解"能力。理解这一点在实践上有直接用处：模型的输出本质是概率取样，所以同一问题可能得到不同答案（可调温度控制随机性），也可能生成流畅但错误的内容（幻觉）。

❓ 自测 2：用"注意力 = 加权平均"解释：为什么模型能分清"小月把画架搬到窗边，因为它光线最好"里的"它"指窗边？

答案：处理"它"这个词时，模型计算"它"与上下文每个词的关联分数（查询与键的点积），"光线""窗边"得分高，经 softmax 变成权重后，对"窗边"的权重远大于"画架"；随后按权重对上下文做加权平均，于是"它"的向量里主要融入了"窗边"的信息。这些权重不是人写的，是训练中从海量文本学出来的——模型见过大量类似的指代现象，总结出了"看语义搭配决定注意力"的规律。

❓ 自测 3：一位同学说"AI 说关羽用过青龙偃月刀重八十二斤，我直接写进解说词了"。问题出在哪？应如何处理？

答案：这是典型的"事实类输出未核查"。青龙偃月刀与重量出自《三国演义》（小说），与史实记载不同，模型往往把小说、影视与历史混为一谈。处理方式：① 判定这句话属于"查得到的事实"（历史细节），必须核查；② 用核查三板斧——要求模型给出处（它可能承认"出自小说"）、换一种问法交叉验证、用联网搜索/权威文献核对；③ 无法核实的表达改为模糊化（"相传他持一柄长刀"）或删除。涉及文化内容的准确性，在第 2、7 章还会反复强调。

❓ 自测 3b：语音识别（STT）与语音合成（TTS）都属于"处理声音的 AI"，为什么本书把它们算作两族不同的技术？

答案：因为任务方向与模型类别都相反。TTS 是**生成**：给定文本造一段语音，同一个句子可以有无数种合格的读法（一对多），模型的容错空间大；STT 是**判别**：给定音频找出唯一正确的文本（多对一），声学证据本身高度歧义（同音词、口音、噪声），模型必须消歧。数据、模型结构、评估指标、失败模式全都不同：TTS 看自然度与相似度，STT 看词错率与同音字错误。这也正是第 1 章开头"识别 / 生成 / 决策"三类分工在声音上的具体体现——**把不同族的技术分清，才知道该用谁的参数、谁的评估标准。**

❓ 自测 4a：什么是"上下文窗口"？为什么说"模型没有记忆"？这两个说法矛盾吗？

答案：上下文窗口是模型一次能读入的词元上限，相当于它的"工作台"大小。说"没有记忆"，指模型不保存跨会话的信息——每次回答时它"知道"的只有：①训练时固化进参数的知识（有截止日期、无法更新）；②本次会话里你提供的全部内容（对话历史、粘贴的文档等）。两个说法不矛盾：窗口内的一切都是"你给它的"，第 8 章讲"记忆机制"时，做的正是"把该记住的东西自动塞进窗口"这件事（记忆 = 外部的存储 + 按需拼进上下文）。实用推论：跨对话框继续工作时，要主动把关键结论重新贴进去；重要指令放开头或结尾，并重复一次。

❓ 自测 4b：一位同学说"我用了温度 0.2，所以每次结果都一样，很可靠"。这句话有什么问题？

答案：低温度让输出更稳定，但"更稳定"不等于"可复现"：只要工具没有固定随机种子（seed），取样仍有随机性，多次运行可能出现细节差异；即便固定了种子，模型的版本更新、后端负载变化也可能导致结果漂移。更重要的逻辑问题是：输出的"可靠性"取决于正确性而非稳定性——温度 0.2 也可能稳定地给出一个错误答案（比如稳定地把"扁舟"标音标错）。正确的心态：低温度用于"要格式稳定"的任务，正确性永远交给核查。

❓ 自测 4：把下列需求补全为"四要素提示词"：给小组的短片生成一段 20 秒的配乐。

答案示例：（角色）你是一位影视配乐师；（任务）为古诗词短片《春江花月夜》的"月升"段落生成 20 秒纯音乐；（约束）中国民乐风格，古琴/箫为主奏，情绪由静到微明，不要人声，结尾自然收束（不要突然截断）；（输出格式）生成 3 个候选版本，每版附一行"情绪与配器说明"，并标注时长。（加分项：参考素材——可附上分镜图或氛围关键词，如"月色、潮水、空旷"。）注意时长必须明确到秒——"20 秒的自然收尾"是传统曲库很难给、而生成工具最擅长的需求。

❓ 自测 4c：小舟的疑问："既然模型每步都是按概率取样，为什么有时它给出的答案又非常固定？"怎么解释？

答案：因为概率分布的形状本身是有偏向的：当上下文把某个答案的概率推到很高（如"1+1="后面的分布几乎全部集中在"2"），即便取样也是"稳定"的。常见于三类情形：①答案是常识或强模式（算术、固定格式、名言）；②温度很低（分布被压尖）；③后训练让模型在特定任务上"高度自信"（如拒答越界问题时）。反过来，"稳定"不等于"正确"——模型可以稳定地输出同一个错误（如稳定的错误引文），这就是核查不可省略的原因。

❓ 自测 5：为什么本书要求"涉及真人的声音与影像素材，先问授权"？请从法律与职业两个角度作答。

答案：法律角度：民法典对自然人声音的保护参照肖像权规定，未经许可克隆、使用他人声音可能构成侵权；《深度合成管理规定》与《人工智能生成合成内容标识办法》还要求 AI 合成内容显著标识、不得用于伪造。职业角度：创作者的声誉建立在"可信"上——一次"未授权声音"引发的争议，足以毁掉一部用心完成的作品；而在商业项目中，素材授权链是否清晰，直接决定作品能否交付与商用。因此"先授权、再创作、后标识"应当成为肌肉记忆。

提示词迭代的可视化：一张"返工地图"

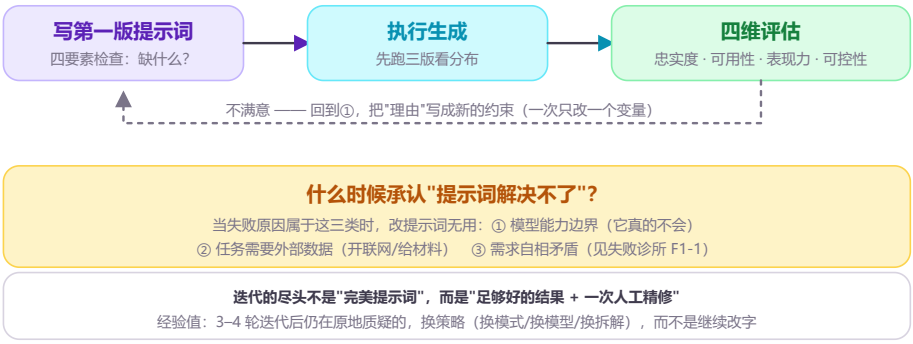


图 1-17　提示词迭代的"返工地图"。把迭代理解为一个有出口的循环：评估给理由、理由变约束、一次改一个变量；判断"提示词解决不了"的三类信号，避免无限返工。

本章方法速查卡：文本与声音创作的"五步"落地

表 1-22　"五步工作法"在本章的落地清单

步骤	文本（脚本）	声音（配音/配乐）	自检问题
① 提出需求	四要素 + 参考资料 + 输出格式（表格列名即接口）	用途、情绪走向、乐器、时长到秒、禁忌项	需求里有没有自相矛盾？占位符都替换了吗？
② 生成	先跑三版看分布；需要规划时开深度思考	分段生成；一次只改一个变量	我是在"批量取样"还是在"期待一次命中"？
③ 评估	四维打分写理由（忠实/可用/表现/可控）	三维评价（自然度/表现力/相似度）	理由是否具体到"下一轮改什么"？
④ 人工修正	核引文、改节奏、定取舍；模型给候选、人来拍板	改标点断句、调语速、拼接分段、配平响度	事实核实了吗？"人工改了什么"记下来了吗？
⑤ 发布/归档	标注 AI 参与情况；存档提示词与参数	加 AI 生成标识；确认音色授权来源合法	三样东西齐了吗：成品、参数、对照结果？

想继续深入？三条学习路径

- **路径一 · 技术原理**：从本课程指定教材《深度学习》第 1 章读起，配合"注意力机制""扩散模型"的公开课程视频；重点关注"为什么这样设计"，而不是推导细节。**适合想转技术方向的同学。**
- **路径二 · 创作实践**：以"每周完成一个小作品"为节奏，从文案、配音、配乐起步，逐步接入后续章节的图像、视频、三维工具；每一次都保存参数记录。**适合想积累作品集的同学。**
- **路径三 · 行业观察**：订阅若干 AI 创作领域的行业媒体与创作者社区，每周记录"一个新工具、一个新案例、一个新争议"；一学期下来你会拥有一份自己的行业观察笔记。**适合关心趋势与就业的同学。**

三条路径不互斥。本课程的任务是把你"通车"到三条路的路口——真正的成长发生在课程结束之后，而它从你保存第一份参数记录时就已经开始了。

把合规变成操作：一份"三查"清单

法规条文读起来遥远，落到操作上其实只有"三查"。把这张清单贴在创作档案的第一页，每次发布前跑一遍：

1. **查素材来源。**每个素材（图、音、字）在档案里都有来源行：来源 URL、授权方式（自有/公共领域/授权素材库/生成）、日期。来源缺一行的素材，不进作品。
2. **查真人权利。**作品里出现真人（本人、同学、公众人物）的肖像、声音、身体影像吗？出现即问：有授权吗？授权范围覆盖本次发布吗？——没有就替换或删除；有则把授权记录存入档案。

3 **查生成标识。**作品包含 AI 生成/合成内容吗（文本改写不算显著、图像/音频/视频通常算）？按《标识办法》添加显式标识（作品内可见的提示或在标题/简介注明），并保留隐式标识（生成工具的元数据不要清理）。**课程所有对外发布的作业一律执行此条。**

这套"三查"与平台的具体操作（各工具的标识开关在哪儿、元数据怎么保留）会随工具更新，附录 E 会给出操作层面的对照表——但清单本身不会过时，它是你在整个创作生涯里的"发布前仪式"。

◆ 拓展阅读

- Ian Goodfellow 等《深度学习》第 1 章与第 10 章（序列建模）——本课程指定教材；配合"深入一格"的注意力公式阅读，理解从 RNN 到 Transformer 的动机。
- 国务院《关于深入实施"人工智能+"行动的意见》（2025 年）与《国家人工智能产业综合标准化体系建设指南》——了解你未来行业"AI 化"的政策坐标，也为本书导论提到的"位置问题"提供注脚。
- 《人工智能生成合成内容标识办法》全文（2025 年 9 月 1 日施行）与配套强制性国家标准——建议通读一遍"标识要求"部分，并在你的第一次作品发布前逐条对照（附录 E 有操作清单）。
- 一本讲"注意力"的科普或视频教程（搜"Transformer 直觉理解"）——各家图解视角不同，多看两三个版本有助于建立直觉。
- 张若虚《春江花月夜》全诗与历代评注——贯穿案例的原典，建议全文背诵前八句（配音脚本会反复用到）。

第 2 章

图像生成与编辑

一句话怎么变成一张画：扩散模型的原理，与网页工具里的参数实操。

学习目标

1. **知识：**理解从 GAN 到扩散、再到流匹配的技术演化；说清"加噪—去噪""潜空间压缩""U-Net 与 DiT""无分类器引导（CFG）"这四个核心机制各自在解决什么问题。
2. **能力：**能读懂并调控采样步数、引导强度、种子、分辨率与宽高比、参考图权重等参数；能独立完成文生图、图生图、局部重绘与条件控制四类操作；能设计并执行"一次只改一个变量"的参数对照实验。
3. **素质：**建立"素材来源合法、生成过程留痕、发布内容标识"的创作习惯；能识别 AI 图像的典型破绽，理解以中华优秀传统文化为题材时的表达分寸与符号敬畏。

对应第 3 周 · 3 学时　讲授 2 学时 · 实践 1 学时　主线推进：角色设定 + 分镜关键帧

小月把第一版"月升"图发到小组群里，三分钟后小舟回了一句："月亮挺好的，就是江面上多了一艘船——我们这一镜里没有船。"小月翻回自己的提示词：**月夜，江面如镜，一轮明月从潮水中升起。**她没写"船"，也没写"不要船"。这正是图像生成最难也最迷人的地方：你写下的每一个词都是一条指令，而你没有写下的部分，模型会按自己的"经验"替你补上。这一章要回答三个问题：一张图究竟是怎么从一串随机数字里"长"出来的（原理）；采样步数、引导强度、种子这些旋钮各自在管什么（参数）；怎么用国内可直接访问的网页工具，把文字分镜一步步变成能进分镜表的关键帧（实操）。学完这一章，你将第一次拥有"把想象变成画面"的可控流程——而不是一次次按下按钮碰运气。

2.1 从 GAN 到扩散：一次范式转移

"让机器画一张不存在的照片"这件事，在深度学习时代被认真攻过两次。第一次的答案是**生成对抗网络**（generative adversarial network，GAN），第二次的答案是**扩散模型**（diffusion model）。理解这两次尝试的分野，你就理解了今天所有图像生成工具为什么长成这样。

◆ GAN：让造假者和鉴定师互相较劲

GAN 的思路非常"聪明"，聪明到近乎狡猾：与其教一台机器"怎么画好一张图"，不如让两台机器互相为难。一台叫**生成器**（generator），负责从随机噪声里造图上；另一台叫**判别器**（discriminator），负责判断"这张图是真的还是生成的"。生成器的目标是骗过判别器，判别器的目标是看穿生成器——两者同时训练，一起进步。

生活类比很直接：**一个造假币的人和一个验钞员互相较劲。**造假者每造一版，验钞员就找出一处破绽；验钞员变强了，造假者就得造得更像。理想情况下，到最后造假者造出的钞票连验钞员也分不出真假——那就是一个"会画图"的模型。

这条路走出过惊艳的成果，但也带着两个著名的毛病。**第一是训练不稳定**：两台机器的进步速度必须大致平衡——判别器太强，生成器得不到有效信号，学不动；生成器太强，判别器又给不出有意义的反馈。训练过程常常像走钢丝。**第二是模式坍塌**（mode collapse）：生成器可能发现"只画少数几类图就能骗过判别器"，于是反复输出高度相似的画面，把训练数据里丰富的多样性丢掉了。用小月的话说："它学会了画月亮，但只会画那一种月亮。"

◆ 扩散：把一道难题拆成几百道小题

扩散模型的思路几乎是反着来的：它不要求网络"一步到位地画出好图"，而是先教网络"如何破坏一张图"，再教它"如何把破坏一步步复原"。

做法分两半。**前向过程**很简单，不需要学习：拿一张真实图片，逐步往里加噪声，加到最后，图片变成一团纯噪声——这一步就像把照片反复复印、直到只剩雪花点。**反向过程**才是要学的部分：训练一个网络，输入"某个噪声程度的图"，让它预测"这一步里被加进去的噪声长什么样"。学会之后，只要反复从图里减掉网络预测的噪声，一团纯噪声就能一步步"显影"成一张清晰的图。

为什么这条路更稳？**因为它把"生成一张图"这个巨大的难题，拆成了几百个"减掉一点噪声"的小题。**每一步的目标都很小、很具体、有明确答案（训练时我们知道自己加了多少噪声，所以可以比对、可以纠正），网络不需要像 GAN 那样在一场零和博弈里小心翼翼地保持平衡，也不需要一次性掌握整张图的全部结构。稳定性的代价是速度：一步到位变成了几十步、上百步的累积计算——这也是后面整整一节要讲的"加速"技术的来由。

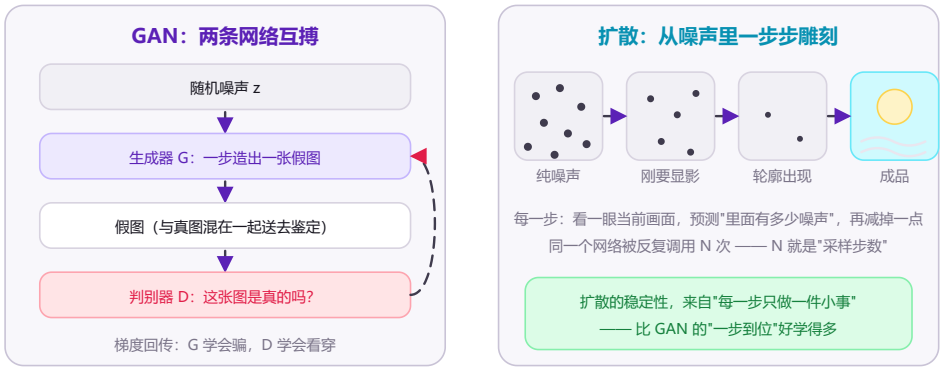


图 2-1　两条技术路线。GAN 用"造假者与鉴定师"的对抗一步出图，训练不稳、容易模式坍塌；扩散把生成拆成"反复减掉一点噪声"的小步骤，代价是慢——于是"加速"成了它必须解决的功课。

◆ 四代骨架：扩散模型的演化地图

扩散模型不是一步长成今天这样的。从它诞生到今天，骨架换了四次，每一次都在回答同一个问题：**怎么让这个"反复去噪"的过程更省、更强、更听话？**

表 2-1　图像生成的"四代骨架"（本书编写时的技术格局）

代际	结构	关键机制	代表
第一代	像素空间扩散	直接在像素上做 DDPM 式加噪/去噪，逐步回归	DDPM
第二代	潜空间扩散（LDM）	先把图像压缩 8×，在压缩后的"潜空间"里做扩散，计算量约下降 64 倍	Stable Diffusion 1.5 / SDXL
第三代	DiT（扩散 Transformer）	用纯视觉 Transformer 处理"切成块"的潜变量，用 AdaLN 注入条件，具备可预测的算力—质量缩放律	DiT（Peebles & Xie, 2023）
第三代变体	MM-DiT（多模态双流）	文本流与图像流各自独立权重 + 双向联合注意力，显著提升文本理解与画面排版	SD3 家族 / FLUX 家族

这张表是本章的"骨架图"：2.3 节讲第二代为什么是分水岭（潜空间），2.4 节讲第三代在架构上换掉了什么（U-Net → DiT），2.5 节讲文本条件怎么被"注进去"，2.6 节讲 2024 年之后为什么整个行业都换了训练目标（流匹配）。

DiT 之后，2026 年前沿出现了三条分叉，值得你留个印象——它们说明"最优架构"这个问题至今没有标准答案：**一是回退到单流**，例如 Z-Image 的 S3-DiT 把双流合并回一条注意力路径，理由是上游的视觉语言模型已经足够强，"模态专属权重"变得多余；**二是混合堆叠**，底层保留 MM-DiT、上层切成单流 DiT（如 LongCat-Image 采用约 10:20 的比例）；**三是稀疏专家（MoE）**，腾讯 HunyuanImage 3.0 采用 64 个专家、总参数 80B、单个词元只激活 13B 的稀疏结构。不过要提醒一句：Qwen-Image、Z-Image、Seedream 等多数前沿系统仍然是稠密模型，"MoE 对稠密"至今没有定论。

> **📌 一个开放问题（2026）**
>
> 检索到的综述把 2026 年图像生成架构最核心的分歧总结为：**自回归式统一（把图像当作词元序列来生成）更容易做强化学习，但更难压缩到少步；DiT 式统一则相反**——前者"聪明但慢"，后者"快但不好用奖励去精修"。这被描述为"2026 年图像生成架构中最具后果的未解问题"。一个佐证是：约 60% 的近期前沿报告（Wan-Image、LongCat-Next 等）已经用**同一个骨干**同时承载文生图、图生图和指令编辑。**注意这是综述观点而非定论**——教材把它放在这里，是让你知道"技术地图的边缘长什么样"，而不是要你记住结论。

还有一件事值得先交代：**训练目标的迁移。**2024 年之后，主流开源模型几乎都从"经典扩散"转向了**流匹配**（flow matching）与其最主流的实现**整流流**（rectified flow）。这是本章最需要写准的一条技术叙事，2.6 节会专门讲。先把结论放在这里：这不是"扩散被淘汰"，而是扩散的血脉延续——换掉的是"从噪声到数据走什么路径"这个设计选择。

表 2-2　GAN 与扩散模型：一张对照表

对比项	GAN	扩散模型（含流匹配）
核心思路	生成器与判别器对抗，一步出图	先学"怎么加噪"，再学"怎么一步步去噪"
训练稳定性	差：两方需保持平衡，易震荡	好：每一步都有明确的学习目标
多样性	易模式坍塌，反复输出相似结果	多样性更好；但也受引导强度影响（见 2.5）
生成速度	快（一次前向）	慢（需多次迭代），故发展出大量蒸馏加速技术
可控性	难：难以精确约束结构	强：条件可注入每一步，利于控制图、参考图、编辑类任务
今天的地位	仍在特定场景使用，已非主流	主流；并进一步迁移到流匹配范式

2.2 加噪与去噪：从噪声里"雕刻"出一张画

现在我们把这台机器拆开。扩散模型的一切，都可以用一句话概括：训练 = 学会去噪；生成 = 从噪声开始，一步步去噪。这一节把这两件事分别讲透。

◆ 前向过程：把一张图"弄脏"，不需要学

拿一张真实的图，记作 x_0。给它加一点点噪声，得到 x_1；再加一点，得到 x_2……这样重复几十到上千次，最后得到的 x_T 已经和原图毫无关系，看起来就是一团电视雪花。这个过程叫**前向过程**，也叫加噪过程。

它最重要的特点是：**完全不需要学习、可精确复现、且答案已知。**因为我们自己决定"在第 t 步加多少噪声"，所以每一步的"标准答案"都攥在手里。这一点至关重要——正是它让后面的训练变成了一件有明确对错的事。

为什么非要"逐步"加噪，而不是一步加到位？因为**从"轻微模糊"到"更模糊一点"是一个小且连续的变化，而"从清晰到纯噪声"是一个巨大的跳跃。**小变化容易学，大跳跃学不会。这就是扩散模型把难题拆小的第一步。

◆ 反向过程：训练一个"噪声侦探"

接下来训练一个网络（记为 ε_θ，意思是"带参数的噪声预测器"）。训练时的流程是这样的：随机挑一张训练图、随机挑一个步数 t、按前面的规则给它加上噪声得到 x_t，然后**把 x_t 和 t 一起喂给网络，问它："这张图里被加进去的噪声长什么样？"**网络给出一个猜测，我们把这个猜测和"我们实际加进去的噪声"逐像素比对，差别越大就惩罚越多，然后稍稍修改网络参数。

重复几百万次之后，网络就成了一个相当可靠的"噪声侦探"：给它一张带噪的图，它能相当准确地指出"这里面有哪一层噪声"。

然后到了**生成**的时候，我们把侦探反着用。手上只有一团纯噪声，没有任何答案。我们让网络看一眼这团噪声，它预测"这里面有噪声，大概是这样"；我们把预测到的噪声减掉一点点，得到一张"稍微不那么吵"的图；再把这张图交给网络，让它重新判断、再减掉一点……这样重复 N 次，画面就一点点"显影"出来。注意：整个过程用的**始终是同一个网络、同一套参数**，只不过被反复调用了 N 次——这个 N，就是你在界面上看到的"采样步数"。

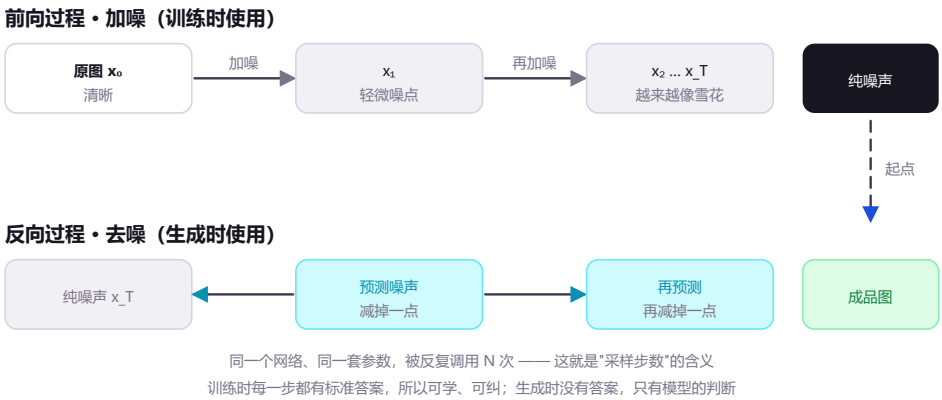


图 2-2　加噪与去噪。上排是训练时用的前向过程（往图里加噪声，答案已知）；下排是生成时用的反向过程（从纯噪声出发，反复"预测噪声、减掉一点"）。两条过程方向相反，共用同一个网络。

还有一件必须讲清楚的事：训练与生成，是**同一个网络的两种用法**，但它们面对的处境完全不同。训练时我们知道答案，所以能算误差、能改进；生成时不知道答案，只能依赖网络"见多识广"后的判断。这就是为什么**模型的训练数据决定了它见过什么、能画好什么**——它没见过的东西，它只能猜。

图 2-3　训练与生成的分工。左边是训练循环（有标准答案，可以纠错），右边是生成循环（没有答案，只能依赖模型的判断）。理解这张图，"采样步数"和"模型能力上限"这两个概念就都落地了。

∫ 深入一格 · 为什么"减掉噪声"就等于"生成图像"　选学 · 数学直觉

你可能会有一个疑问：减掉噪声，凭什么就能得到一张"像样的图"，而不是得到一张"没那么吵的噪声"？

关键在一句话：**网络不是随便预测噪声，而是在"这张图最可能长什么样"的假设下预测噪声。**它是在海量真实图片上训练出来的，见过的图越多，它对"合理画面"的判断就越准。当它看着一团噪声说"这里应该有一轮月亮"时，它实际上是在说：**在所有可能是这张图原貌的候选里，"有一轮月亮"的那个版本概率最高。**

所以每一步"减掉噪声"，本质上是**把画面往"更像真实图片"的方向推一小步。**步数走够了，画面就从"哪个方向都不像"被推到了"某个方向很像"。用一个粗粒度的公式表达这个过程：

$$x_{\text{下一张}} \approx x_{\text{当前}} - \text{网络预测的噪声}$$

这是把 DDPM 的更新式省略掉缩放系数后的**直觉版本**，不是可直接计算的精确公式。它想说明的只有一件事：去噪 = 沿着"更像真实图像"的方向迈一小步。

这也解释了两个常见困惑。**困惑一：模型是在把训练图拼起来吗？**不是。它没有任何"查表"环节，输出的每一个像素都是从噪声算出来的连续数值；这也是为什么它能画出训练集里不存在的构图。**困惑二：为什么同样的提示词每次结果都不同？**因为起点是随机噪声——起点不同，"雕刻"出的成品自然不同。想复现某一次结果，就要固定起点，也就是固定**种子**（seed），这是 2.7 节的内容。

🧪 原理注 · "步数"到底在数什么

训练时的加噪总步数（常记作 T）和生成时的采样步数（常记作 N），**是两个不同的数字。**T 是训练阶段的"最大破坏程度"，可以很大；N 是你实际调用网络的次数，通常只要几十步。少步模型（如 4 步蒸馏）做的事情，正是把 N 压到个位数——代价见 2.7 节。你在界面上看到的"步数"永远是 N。

2.3 潜空间：为什么模型要在"压缩的世界"里画画

第 1 章讲过**潜空间**（latent space）：把信息压缩成抽象表示后所在的那个高维空间。当时我们用文本举例——"月"和"明月"的向量靠得近。在图像这里，潜空间不再只是一个概念，而是一个**实打实的工程部件**：它决定了今天所有主流图像模型的速度上限。

◆ 直接在高清像素上扩散，贵得离谱

算一笔账。一张 1024×1024 的彩色图片有 1024×1024×3 ≈ 314 万个数字。如果扩散过程直接在这么长的数字列表上做，那么**每一步去噪都要让网络读入 300 多万个数、输出 300 多万个数**；而生成一张图要跑几十步。这个计算量会大到让"人人都能用"变成不可能。

更麻烦的是，这 300 多万个数里有大量冗余。相邻像素往往几乎一样，一片天空的两万个像素用"一大片浅蓝"就能描述。**逐像素地去噪，等于在一大堆重复信息上反复做无用功。**

◆ VAE：先压缩，再在压缩后的世界里扩散

第二代扩散模型（潜空间扩散，latent diffusion model）的答案非常直白：**先用一个自编码器把图像"压缩"到潜空间，在潜空间里做扩散，最后再"解压"回像素。**这个自编码器通常叫 **VAE**（变分自编码器，variational autoencoder），它由两半组成：**编码器**把图像压成一小组数字（潜变量），**解码器**把潜变量还原成图像。

典型做法是**空间上压缩 8 倍**：一张 1024×1024 的图，压成 128×128 的潜变量网格。面积小了 64 倍，所以扩散每一步的计算量也大约降到原来的 1/64——这就是"潜空间扩散让 Stable Diffusion 能在消费级显卡上跑"的那条关键机制。生成完成后，解码器再把它还原成 1024×1024 的成品。

用生活类比：**你要修补一条 3 米长的长卷，不是趴在原图上一点点改，而是先把它缩印成一张 A4 小稿，在小稿上修好，再放大回去。**修小稿又快又省力；但代价是——小稿上看不见的细节，放大回去也不会凭空出现。这就是我们要强调的第二点。

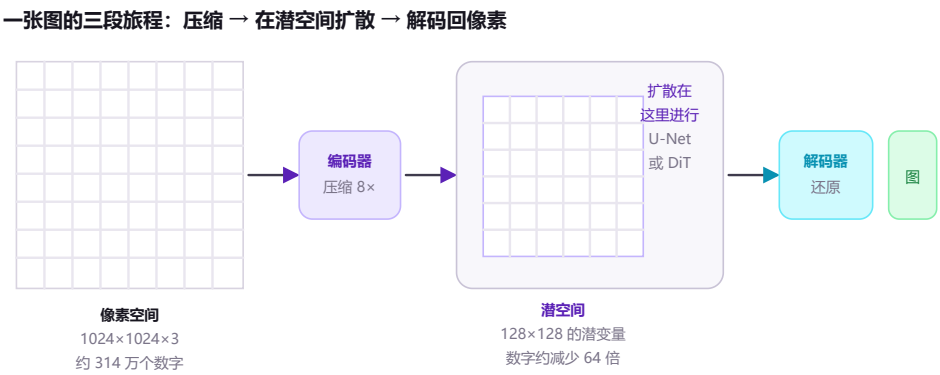


图 2-4　在潜空间里作画。编码器把图像压缩 8 倍（面积缩小 64 倍），扩散模型在这个小得多的表示上反复去噪，最后由解码器还原成像素图。压缩换来速度，也带来细节损失。

◆ 压缩的代价：有损，而且损失是不可逆的

压缩必然有损，这是信息论层面的常识，也解释了一大类你可能遇到过的具体问题：**小字糊成一团、发丝黏连、织物纹理变成噪点、远处的人群像涂鸦。**这些高频、细密的细节，恰恰是"压到 1/64 面积"时最先被牺牲掉的东西。

理解这一点，你就明白两件在实操中非常有用的事。**第一，把"生成文字很密的图"这件事交给图像模型，本身就是逆着它的结构在用力**——它在潜空间里工作，而文字的笔画是最高频的细节。这就是 2.10 节为什么要单独讲"文字渲染"。**第二，压缩率本身也是一个"速度—细节"的权衡旋钮**，行业内一直在尝试不同的压缩方案。

> **🧪 原理注 · 表示空间还在"打架"**
>
> 传统 VAE 的 8× 空间压缩仍是主流，但并非唯一选择。据本书编写时的检索：同一时期并存的方案包括 FLUX-VAE、Qwen-Image 使用的 Wan-2.1 VAE、Wan-Image 自训练的 4 通道 RGBA 设计（**16× 压缩、48 个潜通道**），以及不用 VAE 的 RVQ 分词器。更激进的方向是**像素原生 / 无分词器**：HiDream-O1 于 2026 年 5 月以 MIT 许可开源，去掉了 VAE 和外部文本编码器，其 8B 模型据称在 GenEval 上得 0.90，高于 FLUX.2 Dev（56B）与 Qwen-Image（27B）的 0.87。**该成绩来自厂商/媒体口径，没有独立复核**，教材引用时只当"存在这个方向"的证据，不当结论。

2.4 U-Net 与 DiT：到底是谁在负责去噪

前面反复说"一个网络被反复调用"。这个网络到底是什么？在扩散模型的历史上，它换过一次主角：从 **U-Net** 换成 **DiT**。这次换人不是小修小补，它决定了模型能力"能长到多大"。

◆ U-Net：先退后看整体，再凑近补细节

U-Net 是卷积神经网络的一种结构，名字来自它画出来的形状像字母 U。它的思路分三段：

- **下采样（左边往下走）。**把图像一路缩小——1024 → 512 → 256 → 128……每缩小一次，同样的卷积核就能"看到"更大范围的画面。这解决了卷积网络的一个老毛病：感受野太小，只能看见局部。
- **瓶颈（U 的底部）。**在最小的尺度上，模型处理的是"整张图的全局结构"：构图、大形体、光从哪边来。
- **上采样（右边往上走）。**把特征图一路放大回来，一边放大一边靠**跳跃连接**（skip connection）把左侧浅层的细节"直送"到右侧对应层——这一步非常关键，**因为缩小过程中被丢掉的边缘、纹理，如果不"接力传过去"，放大时就找不回来了**。

生活类比：**画一幅大画时的两种视线。先退后几步看整体布局对不对，再凑近处理笔触和边缘。**U-Net 把这两种视线做进了同一个网络里。它的问题在于：这种"逐层缩小 + 局部卷积"的结构，对长距离关系的建模能力有限——画面左上角的元素和右下角的元素之间要"直接对话"，在 U-Net 里要绕好几层才勉强传得到。

◆ DiT：把图切成"块"，当成一句话来读

2023 年的 DiT（扩散 Transformer）提出了另一种做法，而且思路异常简洁：**既然 Transformer 处理"词元序列"这么强，那为什么不把图像也变成一串词元?**

具体做法是：把潜变量切成一个个小方块（patch），每个小块变成一个向量，于是整张图就变成了一串"视觉词元"——与第 1 章讲的文本词元**结构上同构**。然后交给标准的 Transformer 去处理，每一个块都通过**注意力**与所有其他块直接对话。画面左上角与右下角的元素之间，不再需要绕远路。

条件（时间步和文本）怎么进去？DiT 用的是 **AdaLN**（自适应层归一化）：把"现在走到了第几步""提示词要求什么"编码成一组调制参数，在每一层里直接调节特征的缩放与偏移。**可以粗略地理解成：不是把提示词贴在门口，而是让提示词进入每一间办公室，逐层下达指令。**

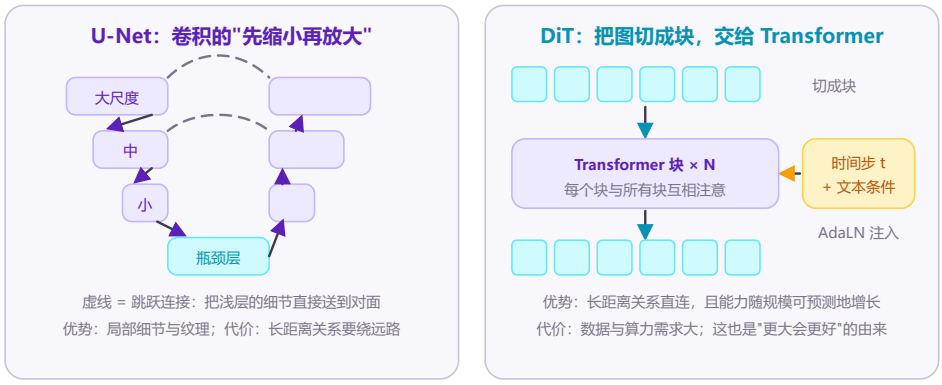


图 2-5　U-Net 与 DiT 的结构直觉。U-Net 靠"缩小看全局、放大补细节 + 跳跃连接"工作；DiT 把潜变量切成视觉词元，交给 Transformer，用 AdaLN 把时间步与文本条件注入每一层。

DiT 最大的收获不是"更准"，而是**可预测**：它的能力随参数量、数据量和算力的增长呈现出稳定的规律（这就是第 1 章讲过的**规模定律**在图像侧的体现）。这让"投更多算力换更好效果"从赌博变成了投资。U-Net 时代，能力靠结构调整抠出来；DiT 时代，能力靠规模长出来——这是 2023 年之后图像模型进步的节奏突然加快的直接原因。

◆ MM-DiT：让文字和画面"各管一段，又互相看得见"

第三代还有一次关键的变体：**MM-DiT**（多模态双流 DiT）。它把文本和图像当作两条并行但有独立权重的流，然后通过**双向联合注意力**让它们互相注视。为什么要这么做?

因为在纯 DiT 里，文本信息是通过 AdaLN 以较"粗"的形式注入的——它能影响整体风格和氛围，但对"画面里到底画了几个什么、文字摆在哪儿"这类精细要求控制力不足。MM-DiT 让文本词元与图像块词元在同一个注意力场里对话，**于是"把标题写在左上角"这种要求，才有可能被真正执行**。SD3 家族与 FLUX 家族是这条路线的代表。

表 2-3　U-Net、DiT、MM-DiT 三者对照

结构	处理单元	条件怎么进来	强项	代表
U-Net	卷积特征图	交叉注意力 + 时间步嵌入	局部细节、纹理、低算力下表现稳	SD1.5 / SDXL
DiT	潜变量小块（视觉词元）	AdaLN 逐层调制	长距离关系、可预测的缩放律	DiT（2023）
MM-DiT	文本流 + 图像流（双流）	双向联合注意力	文本理解与画面排版明显提升	SD3 / FLUX

> **🎯 对你的意义**
>
> 你不需要会写模型，但这两件事能立刻用上。**其一，**看到"这个模型是 DiT / MM-DiT 架构"时，你可以预期它对**文字和版式**的理解更好——这正是你挑选"做海报、做分镜字卡"用哪个模型时的关键信息。**其二，**当你遇到"整体氛围对了、但细部结构总是错"的情况，问题往往不在架构层面，而在**条件控制**（2.8 节）——因为架构决定的是"它能不能理解你的要求"，条件控制决定的是"它有没有被要求必须照做"。

2.5 文本条件与无分类器引导：提示词到底怎么起作用

到这里，去噪的机器已经会转了，但它还只是一台"随机画图机"——你没法告诉它画什么。这一节解决两件事：**提示词是通过什么通道进入去噪过程的**，以及那个最神秘、最容易调错的参数——**引导强度**——究竟在做什么。

◆ 文本怎么"进入"画面：编码 + 注入

第一步是**编码**。你写的中文提示词先被一个**文本编码器**翻译成一串向量——这一步和你在第 1 章学的"切词元 + 嵌入"完全一样。第二步是**注入**：这串文本向量通过注意力机制，在去噪网络的每一层与图像特征发生交互，把"你想要什么"的信息一层层渗进去。

这里有一个容易被忽略的现实：**早期主流文本编码器是在英文语料上预训练的**，高质量的开源标注数据也以英文为主。这直接影响了"中文提示词 vs 英文提示词"的表现差异，2.10 节会专门用一张表来讲这件事。也有专门为中文优化的例子：快手的可图（Kolors）用 ChatGLM3 作文本编码器而不是英文 CLIP，提示词长度约 256 个词元，并构建了覆盖五万余常用汉字的写字数据集，自称是首个原生支持中文文字生成的文生图模型。

◆ 无分类器引导（CFG）：把"提示词带来的方向"放大

现在讲本章最重要的一个参数。**无分类器引导**（classifier-free guidance，CFG），在界面上的名字五花八门——"引导强度""提示词引导系数""提示词相关性""guidance scale"，说的都是它。

它的机制可以用一句话概括：把"有条件的预测"与"无条件的预测"之间的差值放大。所谓"有条件的预测"，是网络看着"月夜，江面如镜"这个提示词给出的去噪方向；所谓"无条件的预测"，是网络完全不管提示词时给出的方向。两者的差值，就代表了"提示词带来的那部分影响"。把这个差值乘以权重 w 再叠加回去，就等于在说：**让提示词的作用再强一点。**

$$\varepsilon_{引导后} = \varepsilon_{无条件} + w \cdot (\varepsilon_{有条件} - \varepsilon_{无条件})$$

ε 表示网络对该步噪声的预测。w 就是界面上的引导强度：w = 1 时等于完全不加引导；w 越大，"提示词带来的方向"被放大得越厉害；w 过大时，画面会被推向"符合提示词但不符合真实图像分布"的极端区域。

注意这个公式里一个很容易讲错的地方：**w 并不是"画得多好"的旋钮，它放大的是"偏离常态"的程度。**w 略大于 1 时，画面确实更贴提示词、线条更锐利；w 再大下去，模型为了满足提示词，会把画面推向一种统计学上"不太可能"的状态——表现为饱和度与微对比急剧上升、高光溢出、皮肤出现塑料感、边缘生硬、阴影过重。这不是"画得更准了"，而是"被推得太远了"。

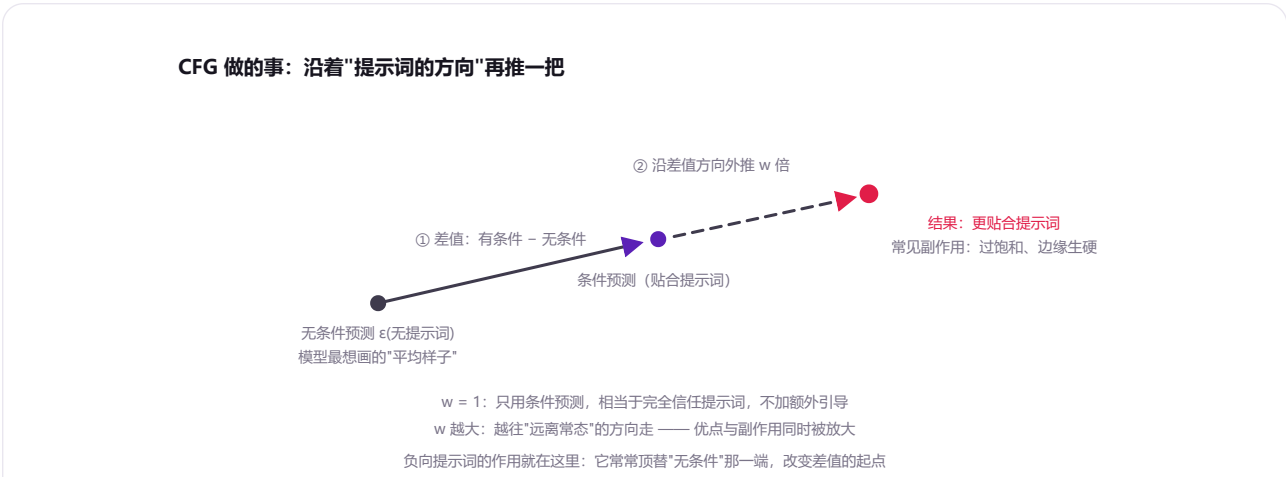


图 2-6　引导强度在做什么。把"有条件预测"与"无条件预测"的差值放大 w 倍，再把结果作为引导方向。w 太小则提示词被忽略，太大则画面被推进"过曝、生硬"的极端区域。

◆ 负向提示词：它到底在"减"什么

很多工具里还有一个输入框叫**负向提示词**（negative prompt）——你不想要的东西写在这里。它的机制，正好是上面公式的一个变体。

在常见实现里，负向提示词**顶替了"无条件"那一端的角色**：模型被要求"从负向提示词所描述的方向上远离"。所以整个引导过程变成了"朝正向走、背向负向走"。理解了这个机制，就能推出三条非常实用的结论：

- **负向提示词里的词，仍然是"被提到的概念"。**模型必须先在某种意义上"知道"这个概念，才能把画面推离它。所以写"不要六根手指"等于把"六根手指"这个概念写进了条件——它可能反而提醒了模型。
- **负向提示词是"缓冲"，不是"补救"。**它能让画面稍微收敛，但效果远不如把引导强度本身调低。遇到"过饱和、高光溢出"这类问题，优先降 CFG，而不是往负向框里堆词。
- **在少步蒸馏模型上，负向提示词基本失效**（原因见 2.7 节）。

◆ 引导强度没有"通用最佳值"

这是本书必须反复强调的一条：**不同模型的引导强度甜点区差异极大**，照抄几年前的老教程（"CFG 设 7"）是学生最常见的错误之一。下表是本书编写时检索到的推荐区间，请把它当作"起点参考"而不是"物理常数"。

表 2-4　常见模型的引导强度参考区间（编写本书时的检索口径）

模型 / 类型	引导强度参考值	备注
SD 系（SD1.5 / SDXL 一类）	7–10	传统扩散的经典区间；但 2025–26 年的新模型普遍低于这一区间
Z-Image-Base	3.5–6（4.5 为甜点）	明显低于 SD 时代的经验值
Z-Image-Turbo	6–8	建议从 7.5 起步；12 以上不推荐
Flux.1-dev	约 3.5	**重要陷阱：**这里的 guidance_scale 是蒸馏出来的 guidance embedding，**不是传统意义的 CFG w**
少步蒸馏模型（LCM / Turbo / Lightning 一类）	1 或 0	必须关闭引导；保持 7 会输出过曝、过饱和的画面

> ⚠️ **CFG 的三条准确表述（可直接背下来）**
>
> ① **CFG 调大 ≠ 图更好。**高 CFG 提升提示词遵守度的同时，会推高饱和度与微对比、造成高光溢出、皮肤塑料感、边缘生硬、阴影过重；SD1.5/SDXL 在 w ≥ 10 时可能出现色彩裁剪、渐变断层、"白色变奶油色"。
> ② **高 CFG 救不了模糊的提示词。**模糊的提示词在 CFG 8 下依然模糊，只是更"吵"。引导强度应当作"微调"，而不是"补救"。
> ③ **CFG 过高会模式坍缩。**约 12 以上易出现结构崩坏、局部过曝、手/脸/文字/机械结构扭曲；21 以上通常构图崩坏，不建议使用。

如果确实需要"高提示词贴合度 + 不炸的画面"，工程上有一些更精细的做法：**CFG-rescale**（用 φ ≈ 0.5–0.7 的系数把引导后的分布重新拉回合理范围）、**APG**（自适应投影引导）、**动态阈值**（dynamic thresholding，把中间步的预测值压在合理区间），以及**只在部分步数区间内开引导**（例如 `guidance_interval=[0.2,0.8]`，只在中间噪声水平阶段施加引导）。这些大多出现在开源工作流（ComfyUI 一类）里，网页工具通常不暴露——但知道它们存在，你就能理解"为什么网页工具里把 CFG 拉到 8 就会过曝，而教程里的参数却很复杂"。

> **🎯 两条"教材级"的表述**
>
> **其一，CFG 与采样步数不是"质量旋钮"，而是"权衡旋钮"。**它们的甜点区，恰恰是"提示词遵守度已经提高，但色彩和光照还没有变得戏剧化"的那个窗口。
>
> **其二，负面提示词只能缓冲，不能替代降低 CFG。**直接调低引导强度，比在负向框里写满"不要过曝、不要过饱和"更管用。

2.6 流匹配：为什么 2024 年后成了主流

如果你只看 2024 年之后的模型发布说明，会发现一个名字反复出现：**流匹配**（flow matching），以及它的搭档**整流流**（rectified flow）。本节把这条技术叙事讲准——因为它是"为什么现在的图出得这么快"的答案。

◆ 先分清两件事：流匹配是框架，整流流是其中一种实现

这是最容易被讲错的一点。流匹配是一个"框架"，它允许你在噪声与数据之间选择任意一条概率路径；整流流则是这个框架里"选择走直线"的一种具体实现，并且通过迭代 reflow（反复重整流）把已经学到的路径进一步拉直，从而减少需要的积分步数。把两者等同起来，就会错过整个要点的关键——"拉直路径"才是省步数的原因。

生活类比：**从山脚到山顶。**经典扩散走的是那条蜿蜒的盘山小路：每一步都要重新判断方向，路很长，必须迈很多小步才不会掉下去。整流流想做的是"尽量走直线"：路径直了，方向变化小，一步就能迈得很大——4 步、甚至 1 步就能到。这就是"少步"从何而来。

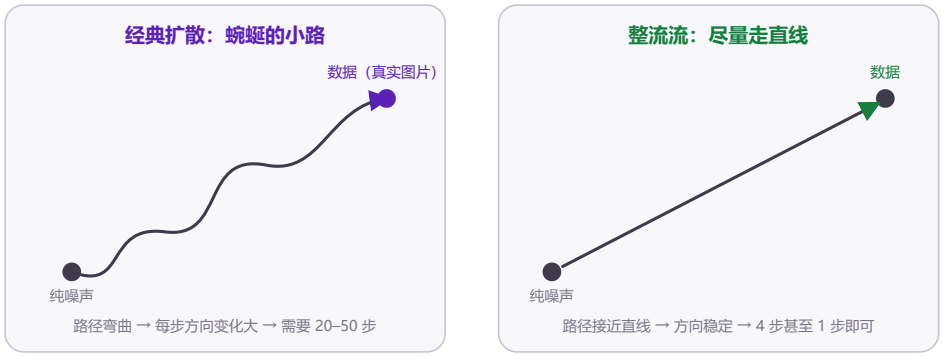


图 2-7　两条路径。左边是经典扩散的弯曲路径（要很多小步），右边是整流流的近直线路径（几步即可）。"步数少"不是魔法，而是路径被拉直的结果。

◆ 迁移的证据：开源阵营几乎集体转向

这条技术路线不是空谈，它在 2024 年之后被开源阵营整体采纳，留下了清晰的时间线：

- **SD3（2024 年 3 月）与 SD3.5（2024 年 10 月）**确立了开源的 Rectified-Flow MM-DiT 基线，规模覆盖 2.5B–8B；其中 **SD3.5 Large Turbo 直接以 4 步蒸馏变体发布。**
- **FLUX.1（2024 年 8 月）**以 **12B** 的 Rectified-Flow 参数成为事实上的开源前沿；**FLUX.1 Kontext（2025 年）**把流匹配推进到图像编辑市场。
- **NVIDIA SANA / SANA-Sprint**让"笔记本上跑 FLUX 级质量"成为可能——靠的正是流匹配使线性注意力、**32× VAE 压缩**与 1–4 步采样同时可行。

学术侧的表述也很直接：CVPR 2026 的一篇论文（FlowSteer）直接写"**流匹配已成为视觉生成的主导范式**"；另一篇安全对齐方向的论文写道"现代扩散模型越来越多采用基于 Transformer 的整流流"（并以 MMDiT 架构的 FLUX 为例），同时指出一个有意思的副作用：**多数安全方法仍针对 U-Net 流程设计，难以泛化到 MM-DiT 上**——技术换骨架的时候，"安全护栏"往往来不及跟上，这是行业里一个真实的缺口。

> ⚠️ **别把它讲成"扩散被淘汰了"**
>
> 准确的表述是：**扩散的血脉延续下来了，具体范式迁移到了流匹配**；整流流是其中最主流的路径实现。而且要注意——**闭源模型（Midjourney、Imagen、GPT Image 一类）并未披露训练目标**，我们只能说"有 Rectified-Flow 式参数化的迹象"，在教材中不应写成确定事实。流匹配也不是扩散的"万能升级"，而是一个有权衡的设计选择。

表 2-5　从扩散到流匹配：变的是什么，不变的是什么

维度	经典扩散	流匹配 / 整流流
从噪声到数据走什么路	预设的加噪曲线，路径弯曲	可自由选择；整流流选"尽量直"的路径，并迭代拉直
典型采样步数	20–50 步常见	少步变可行：4 步蒸馏成常规选项，1–4 步在部分方案中可用
训练信号	预测加进去的噪声	预测"速度场"/路径方向（不同实现细节不同）
不变的东西	仍然是"从噪声出发、迭代逼近数据"，仍然需要文本条件、仍然受引导强度与步数的权衡影响	
对使用者的意义	出图更快更便宜；但"步数必须与蒸馏变体匹配""少步模型下负向提示失效"等新规矩随之而来（见 2.7）	

把这一节压缩一下：**流匹配解决的是"路径太长"这个效率问题，它没有改变"从噪声雕刻出图像"这件事本身。**所以你在 2.2 节建立的那套直觉——训练学去噪、生成靠迭代、步数就是调用次数——依然完全成立，只是现在"调用次数"可以少得多了。

2.7 参数语义：每个旋钮"调大调小"会发生什么

原理讲完了，现在进入本章最重要的实操部分。参数不是玄学——每一个旋钮背后都对应着一条明确的机制，所以"调大调小会发生什么"是可以预测的。这一节把网页工具里你会遇到的参数逐个拆开。开始之前请先记住一条实验纪律：**一次只改一个变量。**同时改引导强度和步数，你永远不知道是哪一项在起作用。

◆ 先补齐"加速"这一课：少步蒸馏模型的三个反直觉

在讲参数之前，必须先交代一个会直接影响你读数的情况：**你在 2.6 节看到的"路径拉直"，并不足以把步数压到个位数。**真正的少步，多半来自另一类技术——**蒸馏**（distillation）：让一个已经训练好的大模型（老师）把自己的"本事"压缩进一个更轻的模型或一层小模块（学生），让学生用极少的步数复现老师的质量。

表 2-6　常见蒸馏加速方法一览

方法	类型	典型步数	交付形态	许可 / 备注
LCM / LCM-LoRA	一致性蒸馏	1–4（常用 2–4）	LoRA（可叠加）	MIT 开源；需配套 LCMScheduler
SDXL-Turbo	对抗蒸馏（ADD）	1–4	完整模型	研究性许可
SDXL-Lightning（字节）	渐进式对抗蒸馏	1 / 2 / 4 / 8	完整模型 + LoRA	OpenRAIL；各步数有独立权重
Hyper-SD（字节）	一致性 + 对抗融合	1–8	统一 LoRA	便于叠加社区模型
Distillation++（ICCV 2025）	推理时蒸馏	4 步 + 1 步教师引导	框架	在 4 步基础上再加 1 步，普遍改善图像质量指标
LogCD（CVPR 2026）	局部 + 全局一致性蒸馏	2–4	统一模型	**全程无需训练图片**，仅约 70 A100 小时

代价必须写清楚。**步数越少，质量越下降**——这不是感觉，是有量化对比的：以 SDXL-Lightning 论文的数据为例，原始 SDXL 在 32 步下的 FID 为 **18.49**；到了 4 步，LCM 为 **21.85**、LCM-LoRA 为 **21.50**、Turbo 为 **22.58**、Lightning 为 **22.30**；8 步时 Lightning 回到 **21.43**（FID 越低越好）。**数字在变差，但速度在变好——这就是权衡。**

产品侧的证据同样清楚：字节 Seedream 4.0 官方称通过**对抗蒸馏 + 分布匹配（10 步以内采样可比肩传统 50 步）+ 4/8-bit 混合量化 + 推测解码与 KV Cache 优化**，实现"推理速度较 3.0 提升超 10 倍""最快秒级生成 2K 图片"；FLUX.2 的 klein 家族（4B/9B）均为 **4 步蒸馏**，官方称可达亚秒级。你在网页

工具上感受到的"怎么这么快"，多半就是这些技术叠出来的。

> ⚠️ **少步模型的"反直觉三件事"**
>
> ① **引导强度必须设为 1 或 0。**把 CFG 保持在 7 去跑少步模型，会得到过曝、过饱和的画面——因为蒸馏过程已经把引导"烘焙"进去了。
> ② **步数必须与蒸馏变体匹配。**4 步的 LoRA 你用 20 步去跑，结果反而更差——因为超出了它被蒸馏出来的那条轨迹。
> ③ **负向提示词基本失效。**同时，提示遵循度会随步数下降、输出的多样性也会变窄。
> 另外：**必须使用模型卡指定的采样器 / 调度器**，换一个可能直接毁掉结果。

◆ 参数语义总表

下面这张表是本章的"参数词典"。请注意表头的前提：**表里的数值都是检索到的推荐值/范围，不是物理常数**；同一参数在不同模型上的甜点区差异很大，实验时务必固定其他变量、只改一个。

表 2-7　图像生成参数语义速查（"参数对照实验"的核心依据）

参数	常见取值	调大 / 调多会怎样	调小 / 调少会怎样
正向提示词 prompt	—	描述越具体，约束越强；但**冲突或过复杂的指令会互相抵消**	越短越自由，模型自行"填空"，氛围容易漂移
负向提示词 negative prompt	—	抑制不想要的元素；但**写"不要六根手指"等于把"六根手指"这个概念写进条件**	少写更干净，尤其对少步蒸馏模型
采样步数 steps	SD 系 20–30；蒸馏模型 1/2/4/8；Qwen-Image-2.1 默认 40；FLUX klein 4 步	细节与收敛更好，但边际收益迅速递减；**超出蒸馏轨迹反而变差**	更快，但提示遵循度与多样性下降；1 步对复杂提示不可靠
引导强度 CFG guidance scale	SD 7–10；Z-Image-Base 3.5–6；Z-Image-Turbo 6–8；Flux.1-dev ≈3.5；蒸馏模型 1 或 0	更贴提示词、线条更锐利；同时**饱和度与微对比急升、高光溢出、皮肤塑料感**；≥12 易结构崩坏	更自由、氛围感强；但忽略提示词、主体不突出、**构图漂移、道具乱跑、手部不稳、数量错误**
种子 seed	整数；如混元 seed 范围 [1, 4294967295]	它不改变"画得多好"，只改变"从哪个起点开始"。**固定种子 = 可复现**（在模型版本、参数、提示词都不变的前提下）	
分辨率 size	即梦 2K 直出、4K 为 AI 增强；Seedream 4.0 默认 2048×2048、最高 4096×4096	细节更多，但**低分辨率下被糊过去的走形会在 2K/4K 全部暴露**；更慢更贵	更快更便宜，但手、文字等精细结构更易崩
宽高比 aspect ratio	常用 1:1、4:3、3:2、16:9、21:9、2.35:1	越偏离训练主分布，越容易出现**元素重复**（"六指的人"、拉长的车）	越接近方形越稳；训练分辨率越单一，非方形越容易出错
参考图权重 --iw / 参考强度	Instant Character 的 scale 默认 1；即梦参考强度需"不高不低"	越贴近参考图，越可能把参考图的构图/杂物一起带进来	越依赖提示词，主体一致性下降
风格强度 --stylize	建议最高 1000；100–400 平衡；500–1000 最大艺术表达	艺术化更强，但越偏离提示词字面内容	更字面、更接近提示词原意
重绘幅度 denoise strength	图生图核心参数	越接近"重新生成"，越可能改动主体身份	越接近原图，改动越弱
控制权重 ControlNet strength	Civitai 建议 0.3–0.7 宽松、0.8–1.2 严格；通义万相/海艺建议 0.6–0.8；Z-Image Turbo 默认 0.6	结构约束更强，但**风格被压平、僵硬、可能出现色带**	结构约束变弱，模型自由发挥更多
分步控制 start_step / end_step	例如 start 0、end 0.5	控制"在哪一段步数区间内施加条件"——只在前半程施加，往往能兼顾结构与自由度	

还有几个参数值得单独说一句。**混乱度**（`--chaos`）：调大后结果更多样、更不可预测，调小则同一批结果更趋同。**怪异度**（`--weird` / `--exp`）：调大更怪诞、更实验。**质量/连贯**（`--q`）：Midjourney V8 的 `--q 4` 能增强复杂场景的连贯性，但**点数消耗变为 4 倍**。这几项的具体取值范围，在所引来源中并未给出，**本书不凭印象填写**——这也是一条给同学的提醒：遇到"范围不明的参数"，正确做法是去官方文档查，而不是照抄二手教程。

A · 模糊提示词　　B · 四要素提示词

图 2-8　提示词详细度的实测对照。同一模型、同一随机种子（20261013），唯一变量是提示词的写法。A 只写「画一只茶盏」，模型自行补全了暗调、无花瓣、无竹影的场景；B 写清了主体、环境、光线与景深，画面元素与提示词几乎一一对应。 AI 生成

C · 采样步数 6　　D · 采样步数 25

图 2-9　采样步数的实测对照。同一提示词、同一随机种子（20261012），唯一变量是采样步数。6 步时花瓣糊成色块、器皿边缘发软、竹影含混；25 步后纹理与边缘才「成形」——这就是「去噪是逐步雕刻」最直接的证据：步数不足，等于雕到一半就停了。 AI 生成

◆ 把参数放在一起看：它们如何共同决定一张图

单个参数看懂了，还要看它们的相互作用。有三组关系最值得记住：

第一组：步数 × 引导强度。这是一对最容易互相掩盖的参数。引导强度太低时，你以为是"步数不够导致画面糊"，其实是"提示词没被听进去"，于是拼命加步数——结果只是更慢地糊着。反过来，引导强度过高时，加步数只会让过曝和生硬更稳定地出现。正确的排查顺序是：先固定一个适中的步数，把引导强

度调到"提示词被听见、画面还没炸"的区间，再微调步数。

第二组：分辨率 × 巧思。2.4 节讲过潜空间的压缩损失。低分辨率下，那些"手指数量不对""衣服上的字是乱码"的问题常常被模糊掩盖；一旦出到 2K/4K，它们会**全部暴露**。所以高清化是"放大器"而不是"修复器"——**不要指望"出个高清版就好了"**。反过来说，做参数实验时先用较低分辨率快速迭代，定稿再出高分辨率，是既省时间又不失真的做法。

第三组：宽高比 × 元素重复。扩散模型在方形图上训练得最充分。当你要求 16:9、21:9 这类非方形画幅时，模型需要"拉伸"它对画面结构的理解，比较容易出现元素重复——同一只鸟出现两次、人的手指多出来、车被拉长。有研究把原因归结为"训练分辨率过于单一"。实操建议：优先选模型预设支持的宽高比，而不是自己硬填一个奇怪的比例。

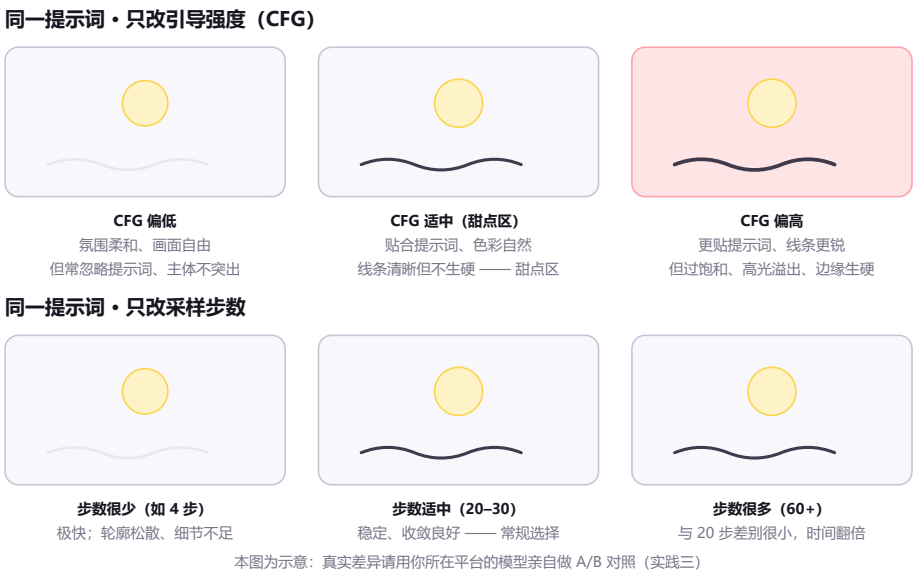


图 2-10　参数效果示意。上排只改引导强度、下排只改采样步数。注意两个"甜点区"的共同特征：**提示词被听见了，但色彩与光照还没有变得戏剧化**。

> 🧪 **原理注 · 种子到底能保证什么**
>
> 固定种子，能让"起点"一样，所以在模型、提示词、参数、分辨率都不变的前提下，结果应当可以复现。但要注意三个前提：① **平台实现可能变**——模型小版本更新、后端优化调整，都可能让同一个种子产出不同的图；② **分辨率或宽高比一变，种子就没意义了**，因为噪声图的形状变了；③ **改蒙版、改参考图，同样会改变输出**，即使种子固定。所以种子的正确用法是"**锁定一次满意结果，便于继续微调**"，而不是"保证永远一样"。

2.8 编辑工具箱：从重绘到条件控制

生成一张图只是开始。真正的创作，八成的功夫在"改"。本节把网页工具里的编辑能力分成六类，逐一讲清楚它们各自解决什么问题、边界在哪里。

图 2-11　编辑工具箱全景。六种能力按"改 / 增删 / 锁定"分成三组。遇到问题先归类，再选工具——这比在界面上乱试快得多。

◆ 图生图与重绘强度：改多少才合适

图生图（image-to-image）做的事情很朴素：不是从纯噪声出发，而是**从你给的图出发**，先加噪声再逐步去噪。加多少噪声、去多少，由一个参数控制——它叫**重绘幅度**（denoise strength，界面上也会叫"重绘强度""参考强度"）。

- **重绘幅度小（例如 0.2–0.3）**：基本保留原图，只改画风、色调、质感。适合"照片转动漫""统一全组镜头的色调"。
- **重绘幅度中（0.4–0.6）**：构图大致保留，但细节会有可观变化。适合"把草图变成成品"。
- **重绘幅度大（0.7 以上）**：接近重新生成，画面会大幅改变，甚至主体身份都变样。适合"只要一个大方向，细节全部重来"。

为什么"放大重绘幅度会改掉身份"？回到 2.2 节的原理就明白了：重绘幅度本质上是"加了多少噪声"。**噪声加得越多，原图残留的信息越少，模型的"自由发挥"就越多。**这是纯机制推出来的结论，不需要死记。

◆ 局部重绘：2026 年最重要的一条分野

局部重绘（inpainting）是"只改我框住的地方"。它可以极大地节省时间，但它的效果好坏事先取决于一件工程事实：**你是用蒙版指定区域，还是只能用文字描述区域。**

据本书编写时对 **36 个云端图像编辑 API** 的第三方核查（2026 年 9 月）：只有 3 个接口接受蒙版文件，而且全部是 GPT Image 端点（GPT Image 1、2.5 Flare、2.5 Sunburst；值得注意的是 **GPT Image 1.5 与 2 反而不支持**）。蒙版用**透明像素**标出应当替换的区域，尺寸必须与源图一致。这组事实的意义在于：**蒙版式与纯提示词式的失败方式完全不同**——蒙版工具允许你"重画边界"来修坏结果；纯提示词工具只能"改写句子"，而且没有可靠办法复现同样的像素。

> **🎯 蒙版的三条实操规则**
>
> ① **蒙版要画得比物体边缘略大一圈**（大约一个软边的宽度）——因为填充是从边界向内生成的，蒙版贴着物体画，会留下"外溢的边"。
> ② **把物体造成的"后果"也框进去。**阴影、反光、被压暗的草地，都是这个物体带来的痕迹；只框物体本身，修完会"痕迹仍在"。
> ③ **复杂主体分两趟。**先修物体，再修阴影/反光。
> 最后记住一条反直觉的事实：**改蒙版会改变输出，即使种子固定。**

◆ 扩图与消除：两条"边界"上的功夫

扩图（outpainting）是向外补画面。它在实操中的作用比想象中大得多：把竖构图补成横构图、把人物周围的留白补出来、把一张正方形图扩成短片需要的 16:9。评测口碑较好的做法之一是 Runway 的 Expand Image，支持扩到原始尺寸的 **2–4 倍**；Canva Magic Expand、Photoshop 的 Firefly Fill & Expand、Clipdrop Uncrop 也都能做。对话式工具里，Gemini 扩得比 ChatGPT 稍多，但质量相近。

消除则是反向操作：抹掉画面里不要的东西。常用工具包括 Cleanup.pictures（去人/物/水印/文字，数秒）、Clipdrop Cleanup（免费版仅支持 720 像素）、Photoshop 的空提示词生成式填充，以及开源的 Lama Cleaner（可本地运行）。

这里有一条必须写进教材的容量边界：**Photoshop 的生成式填充在 8000 像素宽的图上"几乎没用"**；同时，每次生成式填充消耗 1 积分（$19.99 的摄影计划含 1000 积分，约合每次 2 美分）。**也就是说，"用修图软件把大图一点点修好"这条路，在超大画幅下并不可行**——该在生成阶段解决的问题，不要留到后期。

表 2-8　编辑能力 • 典型工具 • 主要边界

能力	英文	典型工具（本书编写时）	主要边界
图生图	`image-to-image`	即梦、通义万相、豆包、Qwen-Image、混元等	重绘强度大时主体身份会被改掉
局部重绘	`inpainting`	即梦"局部重绘 / 控制笔刷"、通义万相、各平台"消除笔"	蒙版支持度差异大；纯提示词式难以复现同一像素
扩图	`outpainting`	Runway Expand Image、Canva Magic Expand、Photoshop Fill & Expand、Clipdrop Uncrop	新区域与原图的光照、透视容易接不上
消除	`cleanup`	Cleanup.pictures、Clipdrop Cleanup、Photoshop 空提示词填充、Lama Cleaner	Clipdrop 免费版仅 720 像素；超大幅面（如 8000 像素宽）下 Photoshop 填充几乎不可用
条件控制	`ControlNet 一类`	即梦（参考图/边缘/深度/姿态）、通义万相"高级控制"、Seedream 原生集成	权重过高会把风格压平
角色一致性	`character consistency`	即梦参考图、Midjourney Edit Model、FLUX.2、Nano Banana 家族、Seedream 4.0	各方案对参考图数量与身份稳定度有明确上限

◆ 条件控制族：让"必须像什么"说了算

这一族技术有一个源头：**ControlNet。**原始论文《Adding Conditional Control to Text-to-Image Diffusion Models》（Lvmin Zhang 等，斯坦福）获得了 **ICCV 2023 的 Marr Prize 最佳论文奖**，代码以 Apache 2.0 开放；它最初围绕 SD 1.5 构建，后来被扩展到 SDXL 与 SD 3.5。

它的核心思想可以用一句话概括：在"提示词负责'想要什么'"之外，再加一条通道，负责"必须像什么"。你给一张结构图（边缘图、深度图、人体骨架图……），模型就被约束着按这张图的骨架去画——内容与风格仍然由提示词决定。

表 2-9　ControlNet 家族常见条件类型

条件类型	英文	提取的是什么	典型用途
边缘控制	`Canny`	干净的对象轮廓线，忽略内部纹理	草图转成品、线框、Logo、产品轮廓；三种主力中**最精确**，但强度过高会压平风格或产生色带
深度控制	`Depth`	每个像素的远近关系	建筑、室内、风景、产品环境图；对纹理颜色宽容，对相机距离与大形体严格
姿态控制	`OpenPose`	人体关键点骨架	动作场景、跨帧一致性、故事板；保留动作意图但**不修复解剖结构**
线稿	`Lineart`	清晰的线条稿（比 Canny 更适合手绘风）	漫画、插画上色
涂鸦	`Scribble`	粗略的涂画线条	快速表达构图意图，灵活度最高
语义分割	`Seg`	按类别着色的区域图（天空/建筑/人…）	大区域布局控制
法线 / 直线 / 细节保持	`Normal / M-LSD / Tile`	表面朝向 / 建筑直线 / 局部细节	三维感增强、建筑校正、放大时的细节保持

三种主力（深度、边缘、姿态）该怎么选？一篇 2026 年 1 月的实测指南给出了相当实用的定位：**Depth 管 3D 空间关系与前后景层次，Canny 管精确轮廓，Pose 管动作意图。**而 IEEE 在 2026 年 6 月的一篇会议论文用它自己的评估指标（SSIM、LPIPS 与 CLIP 文图对齐分）得出的结论是：边缘和深度控制对"结构保持"更好，涂鸦和姿态控制在"灵活性和创意性"上更佳。这两条结论并不矛盾，它们共同说明了一件事——**控制越强，创意空间越小**，这是同一枚硬币的两面。

还有一条非常重要的提醒：Civitai 指出 **2025–26 年的 union 类模型公布的推荐控制权重低于 SD 时代的默认值。**这再次印证了 2.7 节的原则——**不要照抄老教程的参数。**顺带说一句，每种条件类型都是**一份独立的模型文件，必须与基础模型版本匹配**；这也是早年本地使用 ControlNet 时最常踩的坑。

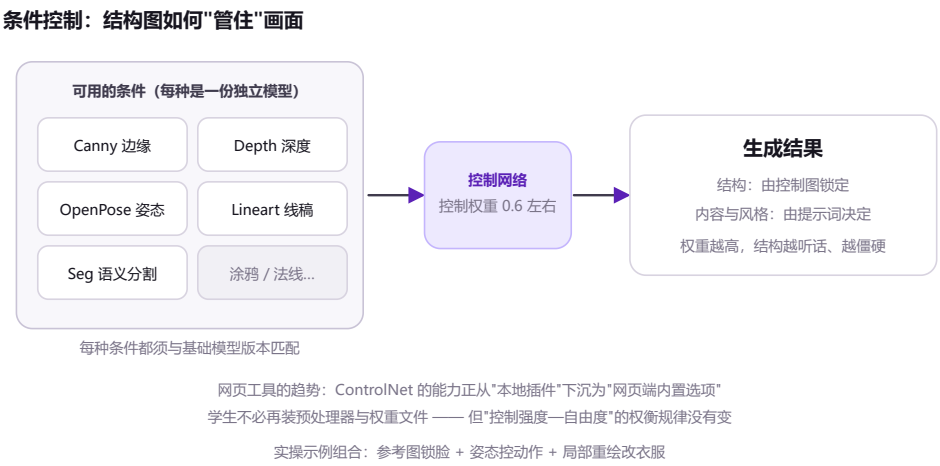


图 2-12　条件控制家族。条件图负责"必须像什么"，提示词负责"想要什么"。控制权重越高，结构越听话，但风格越容易被压平。

表 2-10　网页工具里的"ControlNet 对应物"

网页工具	对应物	说明
即梦	参考图 / 边缘 / 深度 / 姿态 / 局部重绘（四类，可叠加）	官方定位表述为"提示词负责'想要什么'，控制参数负责'必须像什么'"；控制强度"太高僵硬，太低失控"
通义万相网页版	"高级控制"面板内的边缘控制（Canny）与姿态控制（OpenPose）	适合快速验证构图，尤其电商产品图与人物姿态复现
Seedream 4.0	原生集成 Canny、Depth、Mask 等视觉信号，无需额外模型	草图、涂鸦、辅助线即可引导生成，利于姿态控制、建筑设计、UI 原型
云端服务	WaveSpeedAI 的 Z-Image Turbo ControlNet（深度/Canny/姿态，REST 调用，强度可调）；Civitai ControlNet（ComfyUI Cloud，含 35 种预处理器）	无需本地权重与显卡，按调用计费
开源工作流	ComfyUI + 对应节点与权重文件	控制最细（可分步控制），但需要本地环境与显卡

2.9 角色一致性：让同一个人出现在不同画面里

到这一步，你已经能生成、能编辑、能控制结构。但《春江花月夜》的短片有一个绕不过去的坎：**同一个角色，要在 4–6 个不同画面里出现，而且必须看起来是同一个人。**这是生成式图像最难的工程问题之一，也是小月这一周的主要任务。

◆ 为什么"同一个角色"这么难

原因在原理层面很清楚：模型每一次生成，都是从一团新的随机噪声开始的独立采样。它并不"记得"上一张图里那个人的脸。更麻烦的是，多数模型在遇到"多主体"场景时，倾向于采样一张"平均脸"——所以在多人画面里，很容易出现"每个人长得都差不多"的"双胞胎脸"。这个现象有一个真实世界的注脚：2026 年呼和浩特摄影赛一件 AI 作品被识破的证据之一，就是"图中环卫工面容高度相似"。

目前的解法有三条路，按"投入—一致性"排开：

图 2-13　角色一致性的三条路径。A 快而轻，B 慢而稳，C 最轻量、适合原创角色。三者不是互斥关系——实际项目中常常组合使用。

◆ 路径 A：参考图 / 主体参考（零训练）

这是你最先应该尝试的路。做法是：给工具一张或多张"角色参考图"，然后在写提示词时让模型参照它。**参考图数量的上限，是各家工具的硬指标**：FLUX.2 支持最多 10 张参考图；Nano Banana 家族最多 **14 张**参考图（Pro 可稳定相似 5 张人脸、融合 6 张高保真素材，2.1 与 2 稳定保持最多 10 个物体与 4 个人物）；Seedream 4.0 最多十余张参考图，官方称女主角可以在"30 余种迥异的艺术风格与场景"中保持面部特征一致。

通用工具层面，还有一类专门的"角色一致性"服务，例如 fal.ai 的 Instant Character，其默认参数是 `scale=1`、`guidance_scale=3.5`、`num_inference_steps=28`；开源侧则有 InstantID + IP-Adapter（显存要求较低）、PuLID Flux II（无模型污染）等组合。

路径 A 有一条必须写进笔记本的纪律：使用参考图时，提示词只描述情境，不要重复描述参考图里已经有的特征。你写了"戴眼镜的短发女孩"而参考图里那个人就是戴眼镜的短发，这等于**给模型两个"事实来源"**——它需要判断该听谁的，结果往往两边都不像。头发、眼镜、脸型这些"长相"信息，应当完全交给参考图。

◆ 路径 B：LoRA 训练（一致性最高）

如果角色要反复出现很多次，投入一次训练是值得的。**LoRA** 是一种轻量微调技术：用十几到几十张同一角色的图片训练一个小模块，之后生成时只要带上这个模块，模型就"认识"这个角色了。它的工程参数在 2025–2026 年已经相当标准化：

- **素材**：基础角色 LoRA 建议 **15–20 张**自制图；开源社区的做法是 15–30 张多姿势、多表情、多光照的图，并配一个**唯一触发词**（如 `sage_character`）。
- **工程参数**：一份 2026 年的专利申请（公开号 CN122550735A）给出：多视角扩展与数据增强后最终 **20–30 张**（不少于 15 张）；角色 LoRA 的 **rank 取 16–64**、学习率 **1e-4 至 5e-4**、训练 **500–2000 步**；推理时无需再注入参考图。
- **冲突处理**：当角色 LoRA 与风格 LoRA 冲突时，**优先保护面部身份锚点**——脸崩了，风格再对也没用。

这里有一个特别值得学的工程细节：**角色参考表（character sheet）工作流。**做法是用一张照片，经图像模型生成"四列两行"的网格参考图（正面/左侧/右侧/背面各含全身照与头肩裁切）。关键在于——必须使用均匀无阴影的棚拍布光 + 中性灰背景。为什么？因为**均匀的光不会把阴影"烘焙"进素材**；如果参考图里脸上带着一道侧光阴影，下游模型很可能把那道阴影误认成"解剖结构"，在别的光照条件下也执意画出来。一张 1024×1536 的高质量角色参考表，成本约 $0.04116。

◆ 路径 C：精确提示词块与统一角色文件

如果角色是原创的、没有现成参考图，可以先"造"出一个角色，再把它固化成一段反复复用的提示词块——用完全一致的措辞描述同一个人的年龄、发型、服装、体态特征。这条路最轻，但一致性最不稳定，适合"同一批生成里的轻微变化可以接受"的场景。

更工程化的做法是把它变成一个**文件**：ComfyUI 侧的 Omnichar 定义了一种 `.char` 格式，一个文件里同时包含参考图、锁定描述与训练好的 LoRA，可以在图像与视频模型之间通用。**这意味着同一套角色设定，可以同时供给本章（图像）和第 3 章（视频）使用**——这正是小月和小舟在期末之前必须打通的一条管线。

表 2-11　角色一致性三条路径对比

对比项	A · 参考图 / 主体参考	B · LoRA 训练	C · 提示词块 / 角色文件
是否需训练	不需要	需要（约 1 小时）	不需要
一致性强度	中等，受参考图数量与质量影响	最高	较弱，靠文字描述约束
启动成本	约 30 秒级	需要 15–30 张素材与一次训练	最低
典型工具	即梦参考图、Midjourney Edit Model、FLUX.2、Nano Banana、Seedream、Instant Character	ComfyUI 生态、各平台 LoRA 训练功能	复用的角色提示词块、ComfyUI Omnichar 的 .char
最适合	分镜数量不多、需要快速出效果的作业	角色贯穿全片、需要长期稳定	原创角色、暂无参考素材
主要风险	提示词重复描述参考图特征 → 两个"事实来源"打架	素材光照不均匀 → 把阴影烘焙成解剖结构	风格/版本变化导致描述失效

🩺 **案例 2-A** § 小月的角色设定：从"一个模糊的侧影"到一张参考表

背景：《春江花月夜》短片需要一个贯穿全片的角色——"诗人"。小月最初的想法很简单：每张关键帧都写一遍"白袍男子，长发束冠，侧身望月"。结果 5 张图出来，5 个人长得都不一样，其中两张还都是"平均脸"，放在一起像两个人。

做法：小舟建议她按"路径 A → 路径 B"的顺序走。第一步，先用一段稳定的中文描述生成一张"正面、全身、中性灰背景、均匀布光"的角色定妆图，反复微调直到满意，并把这张图定为**参考图**。第二步，用这张参考图为基础，生成"四列两行"的角色参考表（正面/左侧/右侧/背面，各含全身与头肩裁切）——依靠的是"均匀无阴影 + 中性灰背景"这条纪律。第三步，在后续每一张关键帧里，都用这张参考表作为参考图，并且**把提示词里关于长相的描述全部删掉**，只写"月色下他站在江边，衣袂被风扬起"这类情境信息。

结果：重新生成的 5 张关键帧里，人物的脸型、发式、袍服的形制第一次真正"是同一个人"。第一次尝试时小月还在提示词里保留了"长发束冠、白袍"，改掉之后一致性明显提升——这正是"两个事实来源"那条纪律的现场验证。

复盘：①一致性问题的本质是"每次生成都是独立采样"，所以必须从外部提供身份信息；②参考图解决"长得像"，提示词负责"在做什么"，两者不要互相越界；③如果短片篇幅更长、角色出现次数更多，就应该走路径 B 训练一个角色 LoRA，把身份真正"固化"下来。这个过程也直接产出了本章需要交付的"角色设定图"。

2.10 工具地图：2026 年 10 月，你能用上什么

原理讲完了，参数也讲完了，现在必须回答一个最实际的问题：**明天上课打开电脑，我到底该打开哪个网站？**

本章给出的答案是一张**地图**，而不是一份**排行榜**。理由有两层：第一，模型迭代太快——半年就能换一代，任何排名都会迅速过期；第二，工具好不好用，取决于**设备、预算、用途、合规**这四条线，而每个学生的四条线都不一样。记住地图，忘记排名，这是本章最重要的态度。

◆ 选型之前先问三个问题

- **问题一：我要做到什么程度？**课堂练习、海报排版、分镜关键帧——对文字渲染、分辨率、角色一致性的要求完全不同。一张课堂演示图，和一张要放大到两米的展板图，根本不是一件事。
- **问题二：我每个月能花多少钱？**国内工具普遍带有免费额度（例如即梦报道口径为**每天免费 20 次生成**）；国际工具多数**没有免费层**——Midjourney 自 2023 年起取消了免费层。
- **问题三：这张图要拿去做什么？**课堂展示、参加比赛、商业发布，三者的合规要求完全不同。所以你必须回到工具的用户协议原文，而不是相信任何二手转述——**包括本书**。本书编写时检索到国内多数工具的商用条款原文，这条边界必须由你自己动手核实。

图 2-14　图像工具地图。三层结构：国内主线（可直接上手）、国际对照（讲能力与条款差异）、开源本地（能力上限最高、维护成本也最高）。工具名会过时，这层结构不会。

◆ 国内主线：五个"必须认识"的工具

先讲**即梦 AI**（字节跳动·剪映团队）。它是本章教学主线的主角，原因是它的能力覆盖面最完整：文生图、图生图、多图融合、**控制笔刷**（"指哪改哪"）、局部重绘、AI 扩图、局部消除、抠图换背景、风格/姿势保持，还有一个 5.0 时代新增的**检索生图**（实时联网查数据后再出图）。分辨率方面，**2K 直出，4K 为 AI 增强输出**——注意这句话里的分寸："直出"与"增强"是两种不同的东西，考试里会考。

豆包（字节）默认使用 Seedream 系列模型，免费档可用 Seedream 4.0；它的特色功能是**豆包修图**——用语音指令换背景、转风格、创意改图。价格上，官方 2026 年 6 月的声明确认搜索问答、写作生图、语音和视频对话等日常功能**保持免费**，同时推出了专业版三级阶梯（标准 68 元/月、加强 200 元/月、高级 500 元/月），在校大学生认证后有 6 个月特惠（标准套餐 38 元/月）。**对学生的实际含义是：完成本课程的全部作业，用免费档就够。**

通义万相（阿里）是"工具地图"里最需要小心描述的一个。原因见本书的"待核"清单：阿里云百炼官方文档页显示的是 `wan2.5-t2i-preview / wan2.5-i2i-preview`（页面更新至 2026-09-11），而百度百科称 2025 年 12 月已发布万相 2.6 系列，另有来源称百炼已出现 `wan2.7-image-pro` 等型号。多个来源互相矛盾时，教材的正确写法是"以阿里云百炼控制台实际展示为准"——这条写作纪律，比记住任何一个版本号都重要。它真正值得写进笔记的能力是两条：**擅长中英文和小语种文字生成**（支持复杂结构化长文本、图表、架构图），以及网页版在"高级控制"面板内置了**边缘控制（Canny）与姿态控制（OpenPose）**。

Qwen-Image（阿里通义千问）是中文汉字渲染的标杆。Qwen-Image-2.0 于 2026 年 2 月 10 日发布（7B，原生 2K，支持 1K token 提示词），被报道可以"数百字古文全文几乎完全渲染"；2026 年 9 月 20 日开源的 **Qwen-Image-2.1** 采用 7B / 32 层单流 DiT + Qwen3-VL 8B 编码器 + 64 通道 RGBA VAE（16× 压缩），默认 **40 步**，最多 **10 张参考图**，原生支持透明 RGBA。**混元**（腾讯）的 HunyuanImage 3.0 则是全章唯一的"反例式存在"——它**采用统一自回归框架而非主流的 DiT**，64 专家、总参数 80B、激活 13B；它的 Instruct 版本（2026-01-26）把图生图流程做成"先思考，后编辑"。云端价格 **0.2 元/张**（Lite 档 0.099 元/张），V3.0 提供 50 次免费生成（有效期 365 天）。

可图 Kolors（快手）值得记的名字有一串：它是**首个原生支持中文文字生成的文生图模型**，写字数据集覆盖五万余常用汉字；它用 **ChatGLM3** 而不是英文 CLIP 作为文本编码器，提示词长度 256 tokens。它同时也是一堂法律课——见后面的"开源 ≠ 可商用"。最后是 **GLM-Image**（智谱 + 华为，2026 年 1 月开源）：在 CVTG-2K（多区域文字生成准确率）与 LongText-Bench（长文本渲染，覆盖招牌、海报、PPT 等 8 种场景）两项上均获**开源第一**，生成一张图约 0.1 元。

表 2-12　国内主线工具地图（检索口径：2026 年 10 月）

工具（团队）	你要用到它的理由	免费 / 价格要点	必须留意的边界
即梦 AI（字节·剪映）	能力最全的网页工具：文生图、图生图、多图融合、控制笔刷、局部重绘、扩图、消除、风格/姿势保持、检索生图；四类控制可叠加	报道口径：**每天免费 20 次生成**；2K 直出、4K 为 AI 增强输出	平台内版本较多（图片 5.0 Lite / 5.0 Preview、Seedream 4.5 等），以平台实际显示为准；商用条款以官网用户协议为准
豆包（字节）	默认 Seedream 系列，免费档可用 Seedream 4.0；**豆包修图**支持语音指令改图；Seedream 5.0 Pro 于 2026 年 7 月发布	日常功能（搜索问答、写作生图、语音视频对话）保持免费；专业版 68 / 200 / 500 元三档；大学生认证 6 个月特惠 38 元/月	同上，商用条款须以官方协议为准
通义万相（阿里）	擅长中英文和小语种文字生成，支持复杂结构化长文本、图表、架构图；网页版内置 Canny 边缘控制与 OpenPose 姿态控制	百炼新用户约 50 张文生图 + 50 秒视频免费额度（90 天内领用）；文生图约 0.029 美元/张（北京），图生图 0.2 元/张	**版本线混乱**：官方文档与百科、媒体口径不一致，一律以百炼控制台实际展示为准
Qwen-Image（阿里通义千问）	中文汉字渲染突出，可用多种字体渲染数百字古文全文，适合海报、PPT、多格漫画；2.1 支持局部编辑（画圈、涂鸦标注、遮罩）与原生透明 RGBA	2.0 经 API 邀测 / Qwen Chat 免费体验；百炼 `qwen-image-2.0` 0.2 元/张（120 RPM）、`-pro` 0.5 元/张	**Qwen-Image-2.1 采用 Qwen Research License，仅限非商业用途**，商用需单独申请授权；而上一代 1.0 为 Apache 2.0 —— 同名系列内许可差异极大，必须逐版本核对
Z-Image（阿里通义）	高速低价档位（实测 5.1 秒出图）；v2 原生兼容 ControlNet（Canny / OpenPose / Depth / Lineart / IP-Adapter / Tile 六种）	`z-image-turbo` 0.1 元/张（1K）	实测中曾把"2026 年"误渲染为"2016 年"——**快和便宜要付出准确性的代价**
混元（腾讯）	HunyuanImage 3.0 采用**统一自回归框架而非主流 DiT**（64 专家 / 总参 80B / 激活 13B）；Instruct 版本图生图流程为"先思考，后编辑"；覆盖 80 多个训练任务	云端 HY-Image-V3.0 为 0.2 元/张、Lite 档 0.099 元/张；V3.0 提供 50 次免费生成（365 天有效），生成失败不计费	开源权重为 Tencent Hunyuan Community License（第三方对比称月活低于 1 亿可免费商用）；**混元 3.0 参数量有媒体写成 80 亿，与官方模型卡相差 10 倍，以官方为准**
可图 Kolors（快手）	首个原生支持中文文字生成的文生图模型；照片级真实感；用 ChatGLM3 作文本编码器（256 tokens）；提供 ControlNet / Inpainting / IP-Adapter / ID 保持 / LoRA 等完整组件	学术研究免费；商用需先填写快手注册问卷（kwai-kolors@kuaishou.com）	**"代码开源 ≠ 权重可商用"的典型反例**：代码 Apache-2.0，权重为单独模型许可；独立可图平台已于 2024 年 8 月停止提供创作工具
GLM-Image（智谱 + 华为）	2026 年 1 月开源；CVTG-2K 与 LongText-Bench 两项文字渲染评测开源第一	约 0.1 元/张	许可条款未检索到，使用前须核实
文心一格（百度）	仅作历史对照：技术路线"知识增强扩散模型"，240 亿参数	早期实测约不到 4 毛一张	作为独立产品已不存在：2025-02-13 转免费并开放退款，2025-04-01 迁移合并至文心一言，2026 年 6–7 月间相关入口进一步迁移

补充一组**成本对照**（字节 Seedream 系列 API，供你做作业预算时估算）：4.0 为 $0.027/张、4.5 为 $0.04/张、5.0-Lite 为 $0.035/张；5.0 Pro 实测约 0.3 元/张。分辨率口径上，Seedream 4.0 官方称最高 4096×4096、默认 2048×2048，约 1.8 秒出 2K。

◆ 国际对照线：只讲差异，不讲选美

Midjourney V8.2 于 2026 年 7 月 24 日起成为默认版本（谱系：V8.0 Alpha 2026-03-17 → V8.1 2026-04-14 → V8.2 2026-07-24）。它的速度口径是最快比旧版快约 4–5 倍（约 10–15 秒 vs V7 的 30–60 秒），并新增 **Edit Model**，取代了 Omni Reference、Character Reference 与 Retexture 三件旧工具。参数上引入了 `--hd`（原生 2K，无需单独放大）与 `--q 4`（增强复杂场景连贯性，点数消耗 4 倍）。价格体系是本章最有教学价值的一段：Basic

$10/月**起即含商用**，但**年营收超过 100 万美元的公司必须使用 Pro 或 Mega 档位**，且**私有生成需要 Pro/Mega**。**"商用权限"被写进订阅档位、且与公司营收挂钩——这是 AI 工具商业条款的典型形态，值得你在核实国内工具协议时对照着看。**

Adobe Firefly 走的是"多模型超市"路线：平台已托管 30 多个模型（含 GPT-Image 1.5、GPT Image 2、Google Nano Banana 2、FLUX.2 [pro]、Runway Gen-4.5、Kling 3.0 等），自有模型 Firefly Image 5 于 2025 年 10 月 MAX 预览、2026 年 4 月进入 Photoshop 27.6，支持整图变换与风格重绘，最高约 2K（约 4MP）；2026 年 3 月还开启了 Custom Models 公测（用自己的图训练风格/角色一致性）。订阅 Standard $9.99/月（2,000 积分）起。但这里有一条极易被忽略的条款陷阱：Adobe 自有模型定位为"商业安全训练与生产使用"，而第三方合作模型并不继承同等训练与许可保证——你必须逐个追踪最终资产的来源模型。

FLUX.2（Black Forest Labs）是开源/开放权重阵营的代表，也是"许可分层"最细的一家：FLUX.2 [dev] 32B（2025-11-25）为非商用许可；FLUX.2 [klein] 家族（2026-01-15 发布）中，**4B 版本是 Apache 2.0**（约 8GB 显存即可运行），而 9B / 9B KV 为非商用许可；另有约 50 步的 undistilled Base 变体供微调与 LoRA 训练。参考图能力上，klein 最多 4 张，max/pro/flex 通过 API 最多 8 张、playground 最多 10 张。

Google 的 Nano Banana 家族是"对话式编辑"路线的代表：从 Nano Banana（Gemini 2.5 Flash Image）到 Nano Banana 2（`gemini-3.1-flash-image`，2026-02）、Nano Banana Pro（2026-05-29 正式可用）、2.1 与 2 Lite，最多支持 **14 张参考图**（Pro 可稳定相似 5 张人脸、融合 6 张高保真素材），1K/2K 已正式可用、4K 预览（最大 4096×4096），Pro 支持 Google 搜索接地，2 与 2.1 还支持视频输入。**每张输出图都带 SynthID 隐形水印与 C2PA Content Credentials**——把这句和 2.15 节的合规要求放在一起看，你会发现"谁在用机器可读的方式标注 AI"已经是一门显学。与之相对，**OpenAI 的 GPT Image 系**（GPT Image 2、以及第三方清单中出现的 2.5 版本）最强的一点是中文排版能力，以及它是 36 个云端图像编辑 API 中**唯一支持蒙版式局部重绘**的系列（见 2.8 节）。

表 2-13　国际对照线（检索口径：2026 年 10 月）

工具	版本与时间线要点	能力与额度	商用条款要点
Midjourney	V8.2 于 2026-07-24 起为默认；此前 V8.0 Alpha（2026-03-17）、V8.1（2026-04-14）	标准任务比旧版快约 4–5 倍（约 10–15 秒）；`--hd` 原生 2K；`--q 4` 增强复杂场景连贯性；V8.2 新增 Edit Model，取代旧的三件参考工具	**$10 Basic 起即含商用**；**年营收超 100 万美元的公司必须用 Pro/Mega**；私有生成需 Pro/Mega；无免费层
Adobe Firefly	Firefly Image 5：2025-10 MAX 预览、2026-04 进入 Photoshop 27.6；平台托管 30+ 个模型；Custom Models 2026-03 公测	整图变换与风格重绘、自然语言指令；最高约 2K（约 4MP）；Generative Fill 是实用主力	自有模型定位"商业安全训练与生产使用"；**第三方合作模型不继承同等保证**，须追踪最终资产来源模型
FLUX（Black Forest Labs）	FLUX.2 [dev] 32B（2025-11-25）；FLUX.2 [klein]（2026-01-15）：4B 与 9B / 9B KV，4 步蒸馏、亚秒级；undistilled Base 变体约 50 步	文生图 + 单参考/多参考编辑；klein 最多 4 张参考，max/pro/flex API 最多 8 张、playground 最多 10 张；输出最高约 4MP	仅 klein 4B（Apache 2.0）真正无限制；klein 9B 与 dev 32B 非商用——**输出可用于商用，但把模型部署/服务于商业产品需向 BFL 购买许可**
ComfyUI 生态	节点式工作流事实标准：用户超 400 万、社区节点超 6 万个、GitHub 约 10.6 万星；2026-04 完成 3000 万美元融资、估值 5 亿美元	节点式 DAG 工作流；**PNG 内嵌完整 workflow JSON**（拖回即可恢复，实现可审计、可复现）；已支持 SD、SDXL、SD3/3.5、FLUX、Kolors、Z-Image 及多个视频模型	核心开源免费（GPL-3.0）；**代码许可与各模型权重许可彼此独立**，逐个核对；Comfy Cloud 托管约 $20/月
Nano Banana 家族（Google）	Nano Banana → 2（2026-02）→ Pro（2026-05-29 GA）→ 2.1 → 2 Lite	**最多 14 张参考图**；1K/2K 正式可用、4K 预览；对话式编辑、光影/焦距调整、相机变换、品牌一致性；Pro 支持搜索接地；2/2.1 支持视频输入	无免费 API 层；**每张图带 SynthID 隐形水印 + C2PA Content Credentials**；免费层仅存在于 Gemini App / AI Studio
GPT Image 系（OpenAI）	GPT Image 2（2026-04）；GPT-Image 1.5 于 2026-01 上线 Adobe Firefly	中文排版能力强；**36 个云端图像编辑 API 中唯一支持蒙版式局部重绘**的系列（蒙版用透明像素标出替换区域）	完整价目未检索到，使用前须自行核实

◆ "开源 ≠ 可商用"：一张必须单独讲清的清单

上一节的表格里藏着一类特别危险的误解：很多同学看到"开源"两个字，默认它就是"随便用"。事实是：代码的开源许可与模型权重的许可，是两份彼此独立的文件。ComfyUI 的代码是 GPL-3.0，但你在它里面跑的每一个模型权重，都带着自己的一份许可。下面这张清单请直接抄进笔记本。

表 2-14　开源 ≠ 可商用：权重许可清单（检索口径：2026 年 10 月）

模型 / 组件	许可	商用结论
FLUX.2 klein 4B	Apache 2.0	可商用（官方称约 8GB 显存即可运行）
FLUX.2 klein 9B / 9B KV、FLUX.2 [dev] 32B	FLUX 非商用许可	生成**输出**可用于商用；但把**模型本身**部署或作为商业产品的一部分对外服务，需要向 BFL 购买许可
Qwen-Image-2.1	Qwen Research License	**仅限非商业用途**，商用需单独申请授权
Qwen-Image 1.0	Apache 2.0	据第三方对比，无收入或用户数限制 —— 与 2.1 形成鲜明对照
可图 Kolors	代码 Apache-2.0；权重为单独模型许可	学术研究免费；**商用需先填写快手注册问卷**
HunyuanImage 3.0	Tencent Hunyuan Community License	据第三方对比：月活低于 1 亿的组织可免费商用
ComfyUI	GPL-3.0（代码）	代码本身开源免费、无功能阉割；**但其中加载的每个模型权重各按各的许可执行**

> ⚠️ **三个最容易踩的坑**
>
> ① **同名系列内的许可差异极大。**Qwen-Image 1.0 是 Apache 2.0，2.1 却是研究许可；FLUX.2 家族里 klein 4B 是 Apache 2.0，9B 与 dev 32B 都是非商用。**永远逐版本核对，不要按系列名记忆。**
>
> ② **"输出可商用"与"模型可商用"是两回事。**FLUX 的条款里，输出可以用来商用，但把模型部署成对外服务需要另行购买许可。这两句话经常被混为一谈。
>
> ③ **平台条款会变。**平台上的"会员可商用"类说法常见于媒体报道，但本书编写时并未取得国内主要工具官方商用条款的原文。做正式项目前，请自行打开官网用户协议逐条读——**本书转述的任何条款都不能作为法律依据。**

◆ 用中文还是用英文：一个被证明是伪命题的问题

"写提示词应该用中文还是英文？"——2026 年的答案是：这不是选语言的问题，而是选偏差的问题。有跨境实测文章直接指出"哪种更好是伪命题"，真正要理解的是中英文提示词会带来**系统性偏差**。四条证据值得逐条讲：

- ① **底层与 token 效率。**主流扩散模型多基于 CLIP 等**英文预训练**，高质量社区标注样本 90% 以上是英文；英文有天然空格、token 切分边界清晰，中文常被切成 1–3 个字符的碎片。**但请注意：这条"token 效率"的说法在来源中并未给出量化测量，所以正确的表述是"英文的 token 切分通常更规整"，而不是一个具体倍数。**
- ② **术语精度（有实测数字）。**30 组对比中，中文"白底"提示产生的背景偏暖（约 RGB 248–252，约 60% 带轻微灰、30% 有阴影残留）；写成 `studio shot on pure white background (#FFFFFF)` 则达到 RGB 254–255，约 85% 完全干净。亚马逊的商品图标准是"纯白背景，RGB 255,255,255"——在电商场景里，"中文提示词"经常直接意味着返工。
- ③ **风格词颗粒度。**英文视觉风格词经广告/摄影/设计行业标准化，颗粒度细、稳定性高。实测中 `brutalist undertone` 能收紧背景质感与边缘光，而中文对应表述**每 10 张里约 2–3 张发生漂移。**还有一个特别生动的例子：中文"**复古**"指向 1980–90 年代港风，英文"**vintage**"却指向 1950–70 年代美式——同一个词，两个时代。
- ④ **文化适配。**春节场景用中文最好（红梅、窗花、金箔保持完整），翻译成英文会漂向"泛亚洲节庆"；中秋用中文能保持月饼/茶席构图，译成英文会漂向日本茶道；反过来，圣诞用英文最好，译成中文会塌成"商场橱窗装饰 + 塑料感松树"；万圣节偏英文。

基于这些偏差，实测验证过三种**混合策略**，直接可以用在作业里：**组合 A**——中文主体 + 英文风格/技术词（例如明式家具用中文描述、风格部分写 `studio shot, minimalist, soft top light, neutral gray background`），比全中文稳定约 30%；**组合 B**——英文主体 + 中文文化锚点（例如意式咖啡机放在春节场景里），氛围细节比全英文深 50% 以上；**组合 C**——英文参数与比例 + 中文叙事描述，有助于批量一致性。至于**不建议混用**的场景：简单白底图、创意多样性探索、节庆图像（节庆图像用"该文化的母语"最稳）。

另外两个容易忽略的实操坑：**2026 年 4 月的实测显示中文标点识别仍不可靠，建议使用半角标点**；以及 2.9 节已经讲过的——**重复描述参考图已有特征，等于给模型两个"事实来源"。**

表 2-15　中文提示词与英文提示词：差异与混合策略

维度	英文通常更占优	中文通常更占优	实操建议
底层与 token	主流模型基于英文预训练，社区标注样本 90% 以上是英文；token 切分边界清晰	中文优化模型（Kolors、Qwen-Image、Seedream、混元）已显著缩小差距	追主流模型用英文；用中文优化模型时按该模型实测
术语精度	白底达 RGB 254–255、约 85% 完全干净	中文"纯白底"约 RGB 248–252，约 60% 带轻微灰	电商/产品白底图一律用英文写法
风格词	行业标准化、颗粒度细、稳定性高；中文表述每 10 张约 2–3 张漂移	"复古"能准确唤起 1980–90 年代港风	风格词用英文，文化意象用中文
文化适配	圣诞、万圣节	春节（红梅/窗花/金箔完整）、中秋（月饼/茶席）	节庆图像使用该文化母语
混合策略	A：中文主体 + 英文风格/技术词（稳定约 +30%）｜B：英文主体 + 中文文化锚点（氛围细节 +50% 以上）｜C：英文参数比例 + 中文叙事（利于批量一致性）		按作业类型选组合，别凭感觉随时切换
标点与参考图	中文标点识别仍不可靠（用半角标点）；重复描述参考图特征会引入第二个"事实来源"		把这两条写进自己的检查清单

◆ 让图"写字"：文字渲染的真实水平

海报、封面、信息图——数字媒体创作绕不开"画面里的字"。2026 年的状态可以用一句话概括：从"普遍乱码"进步到"主流模型基本可用"，但复杂信息图与长文本仍有明显错误。这句话的两半同样重要：前一半告诉你"可以放心用了"，后一半告诉你"不能不看就交"。证据如下。

表 2-16　文字渲染的真实水平（证据表）

证据	内容	口径说明
百炼三档模型实测（2026-09）	活动海报（10 处文字要素）三档**全部 10/10 全对**；高密度会员权益说明（9 处要素）中两个高配档位两轮全对（19/19）；`z-image-turbo` 虽 5.1 秒出图、0.1 元/张，却把"2026 年"误渲染为"2016 年"	第三方实测；单价：z-image-turbo ¥0.1、qwen-image-3.0 ¥0.18（该型号存在性存疑）、3.0-pro ¥0.25
Qwen-Image-2.0（2026-02-10）	中文汉字渲染突出，"数百字古文全文几乎完全渲染"，适合高级海报、PPT、多格漫画；支持 1K token 超长文字输入	官方与媒体报道口径
GPT-Image-2（2026-04）	SuperCLUE 评测汉字生成 93.07 分、文字准确度满分；实测文章称中文排版准确率约 99%	媒体/厂商口径，非独立基准
GLM-Image（2026-01）	CVTG-2K 与 LongText-Bench（招牌、海报、PPT 等 8 种场景）均获开源第一	官方口径
Seedream 5.0 Pro（2026-07）	中文小字排版不再乱码，与 GPT Image 2 文字质量基本打平；但**复杂中文信息图仍有错字**——一张 AI 模型入门指南网格卡片出现 5 处错误；字体变化性略逊	第三方实测；成本约 0.3 元/张
Nano Banana Pro	一处引用称文字渲染准确率约 64%；另一来源却称 Nano Banana 2 的文字准确率实际**高于**Pro	两处口径张力明显，Google 未公布 2.1 的任何基准——**付费前务必自行 A/B**

把这张表读成一个操作结论：**海报类中文文字渲染已普遍达到可交付水平**（实测全对或接近 99%），差距主要在**字体多样性、复杂信息图稳定性与价格**；而**高密度信息图与长文本的逻辑一致性仍是主要失分点。**所以本章给出的作业建议是——图里的字要"少而准"，大段文字交给后期排版兜底。

2.11 关键概念表

本章新出现的概念比较多，这张表是给你的"随身卡片"：每个概念用一句话解释，并标注它在本章的出处，方便回查。

表 2-17　第 2 章关键概念表

概念	一句话解释	本章出处
GAN（生成对抗网络）	生成器与判别器互相博弈，像"造假币者"对抗"验钞员"	2.1
模式崩溃	生成器发现少数几种输出就能骗过判别器，于是反复只生成那几种	2.1
扩散模型	训练时学会"识别并去除噪声"，生成时从纯噪声出发逐步去噪，把图像"雕刻"出来	2.2
加噪 / 去噪	正向过程按固定规则加噪；反向过程由网络预测噪声并逐步减去	2.2
潜空间	把像素图压缩后的低维表示，是扩散实际发生的"工作台"	2.3
VAE 编解码器	负责把图像压进潜空间、再把潜变量解码回像素图	2.3
U-Net	早期扩散主干：下采样到瓶颈再上采样，靠跳连保留细节	2.4
DiT（扩散 Transformer）	把潜变量切成 patch 当视觉 token，用 Transformer 处理，规模可扩展性更好	2.4
MM-DiT	文本与图像双流并行、双向联合注意力的架构（SD3 / FLUX 系）	2.4
文本条件	把提示词编码后注入去噪网络，决定"生成什么内容"	2.5
CFG（无分类器引导）	用"有条件"与"无条件"两次预测的差值放大提示词影响，w 越大越贴提示词、副作用越多	2.5
负向提示词	"无条件分支"的替代品：写"不要六根手指"等于把"六根手指"写进条件	2.5
流匹配 / 整流流	预测"速度场"、把噪声到图像的路径拉直；框架与实现两个层次要分开说	2.6
蒸馏加速	用少步学生模型模仿多步教师模型，1–8 步出图，代价是多样性与质量下降	2.7
种子（seed）	同一模型、同一参数、同一种子才可能复现同一张图；换模型/分辨率即失效	2.7
重绘幅度	图生图里"加多少噪声"，越大越接近重新生成、越易改掉主体身份	2.8
局部重绘 / 蒙版	只改指定区域；蒙版要略大于目标、包含后果、可重画边界	2.8
ControlNet 条件控制	用边缘/深度/姿态等结构图约束构图，权重越高越听话、风格越平	2.8
角色一致性	让同一角色跨画面保持身份：参考图、LoRA 训练、提示词块三条路径	2.9
LoRA	轻量微调模块，用 15–30 张素材把角色特征"固化"进一个小文件	2.9
C2PA / SynthID	内容凭证与水印机制，让 AI 生成痕迹可溯源——合规与鉴定的技术基础	2.10 / 2.15

2.12 案例教学：从文字分镜到六张关键帧

理论、参数、工具都齐了，现在进入本章的案例教学。这一节有两个案例：**主线案例**展示春江小组如何把第 1 章产出的文字分镜，变成一组风格统一、角色一致的图像关键帧；**外部案例**看真实产业界在 2025–2026 年是怎么做的——以及它们摔过的同一个跟头。

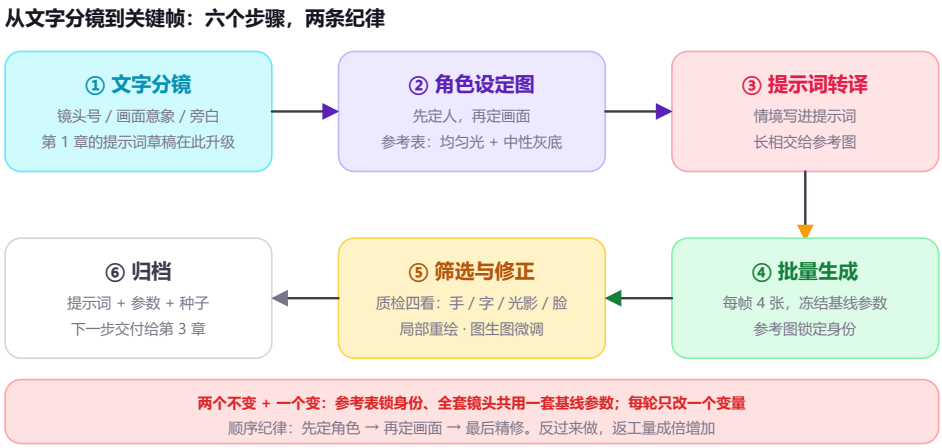


图 2-15　从文字分镜到关键帧的六步工作流。第②步是最容易被跳过、也最不该跳过的一步：角色设定一旦定稿，后面所有画面才有"同一个人的参照系"。

分镜 1 · 空镜　　分镜 2 · 花林

分镜 3 · 楼台　　分镜 4 · 孤舟

图 2-16　四张成片级关键帧（主线案例成品）。按 2.9 节的「角色与风格参考 + 统一提示词块」流程产出的四帧。注意它们共享同一套视觉语言：青蓝色调、月光为唯一主光源、大量留白——这正是「系列图一致性」要练的东西。 AI 生成

◆ 主线案例：春江小组的"关键帧之夜"

先交代任务背景。第 1 章结束时，小组手里已经有了一份"能开工"的资产：**开头 60 秒的解说词**加上一张**4–8 个镜头的分镜表**，每个镜头还带着一条"可执行的 AI 生成提示词草稿"。本章要交付的是：**一张角色设定图** + **4–6 张分镜关键帧**——它们既是本章的作业成果，也是第 3 章（视频生成）的首帧素材。

小月把这项工作叫做"关键帧之夜"。她的做法与图 2-15 的六步一一对应，但有两处细节值得单独展开。

细节一：把镜头表"翻译"成画面任务表。文字分镜是给"讲故事"用的，画面任务表是给"生成"用的——后者必须多出三列：**一致性与控制**（这张图需要锁定什么）、**主要风险**（这张图最容易死在哪）、**判定标准**（什么样才算过）。这个习惯来自第 1 章的"风险备注"技巧，本质是一样的：把"验收条件"写在开工之前，而不是生成之后再凭感觉挑。

表 2-18　六张关键帧的画面任务表（主线案例成品）

关键帧	画面意象（对应诗句）	景别	生成要点	一致性与控制	主要风险
K1 月升 · 潮水	"春江潮水连海平，海上明月共潮生"	远景 · 固定	江面占画面约三分之二，一轮明月贴水升起，有薄雾层次	无人物，不挂参考图；色调锚定青蓝	水面反光在跨帧时最难一致（留给第 3 章处理）
K2 水月 · 诗人	"江畔何人初见月"	中景 · 人物偏右	参考表锁身份；提示词只写情境（月色、江畔、衣袂被风扬起）	参考图必挂；提示词删净长相描述	手部动作——只允许自然下垂，不做复杂手势
K3 花林	"月照花林皆似霰"	全景 · 俯拍	花树成片，月光在叶隙间呈霰点状；人物可极小或省略	与 K2 共用同一套色调基线	花朵糊成噪点或出现重复纹理
K4 楼台	"可怜楼上月徘徊"	中景 · 仰视	楼台轮廓清晰，月悬檐角；建筑直线多	用深度或边缘控制锁建筑结构	栏杆、瓦檐等重复元素拉长变形
K5 孤舟	"谁家今夜扁舟子"	远景 · 平视	江面一叶孤舟，舟上人影小而模糊	人影挂同一参考表，但比例极小、不带五官细节	舟形畸变、人影出现"双头"或多余肢体
K6 归途 · 空镜	"不知江月待何人"	远景 · 有缓慢移动感	收尾空镜：月与江面，画面留白最大	无人物；与 K1 保持"同景别但不同构图"	与 K1 相似度过高，观众以为画面卡住了

细节二：冻结一套"基线参数"。小月在第一帧开工前就写死了一组参数，之后所有镜头、所有轮次都不许动，除非做对照实验：**模型档位固定**（同一个平台、同一个模型版本）、**分辨率固定**（16:9 定稿尺寸）、**引导强度固定在甜点区**（先在 4–7 区间内试两个值，选定一个）、**每帧生成 4 张**。她的理由很"工程"：如果每张图都换参数，你永远不知道"前后不像"是参数的错还是模型的锅——这正是 2.7 节"一次只改一个变量"的纪律在真实项目里的样子。

下面两张提示词卡是小月实际使用的版本，你可以直接抄进自己的作业，只改括号里的占位符。注意第一张卡里关于负向提示词的自注——它是本章 2.13 节的一个伏笔。

✎ 提示词 · 关键帧生成（以 K2"水月"为例）

画面情境：月色下，一位白衣束发的男子侧身立于江畔，衣袂被晚风轻轻扬起；远处花林似霰，江面浮着碎银般的月光。构图留白多，人物偏画面右侧，中景。
风格描述：中国山水画的写意气质，青冷色调，唯一光源是月光，胶片颗粒细腻，不要强烈明暗对比。
负向提示词：多指、多手、文字水印、现代服饰、过饱和、塑料感。
（自注：负向提示词对"多余手指"的抑制有限——写"不要六指"等于把"六指"这个概念写进条件。真正的解法见 2.13 失败诊所。）
长相描述：空白——全部交给角色参考表。

参考图 角色参考表（不重复描述长相）　分辨率 16:9 · 2K 档　采样步数 平台默认　引导强度 甜点区（4–7 内试两个值）　每帧张数 4 张
种子 每组实验记录在案

生成回来之后是"筛选与修正"。小月给自己定了一条硬规矩：**任何一张图进候选之前，必须过"质检四看"**——看手（手指数量、指节、手与物体的边界）、看字（画面里有没有乱码文字）、看光影（光从哪来、有没有两套光）、看脸（五官比例、跨帧是不是同一个人）。这套"四看"不是她发明的，它是 2025–2026 年中国多起摄影赛事识破 AI 作品时实际使用的线索清单（见 2.15 节）。

修不满意的图，按代价从低到高走三步：**①局部重绘**（只改坏掉的那一块）；**②图生图**（挂上原图、把重绘幅度调到 0.3–0.4，保留构图换质感）；**③重新生成**（当问题属于"全局关系错误"时，局部窗口装不下它）。最后一类问题长什么样？典型例子是"手与手臂接不上"——它是全局的透视问题，框多大都修不回来，必须重掷。下面是她的修正提示词卡。

✎ 提示词 · 局部重绘修正（手部）

蒙版画法：完整覆盖出问题的手指与指间粘连处，比手指外缘多留约一圈软边；手造成的阴影与反光一并框进去。
局部提示词：五根手指清晰可数，拇指与其余四指自然分开，手背朝向镜头，指节比例正常，月光自右上方来。
纪律：如果两次重绘之后仍然不对，停止重绘——重绘是"重画边界"的机会，不是无限次数的。改走"重新生成"，或用换手法从同批生成里取一只正确的手。

重绘幅度 ≈ 0.4　种子 固定　判定标准（1:1 视图） 手指可数 · 指节正常 · 掌心朝向与手臂一致

🩺 **案例 2-B** § 关键帧之夜：24 张初稿到 6 张定稿

背景：小月要在一次课后练习（1 学时）加一个晚上，完成 6 张关键帧。她的原始想法是"一张一张做，做到满意为止"。小舟提醒她：这样做的结果几乎可以预见——三张之后，色调开始漂，人物开始变脸，时间全耗在返工上。

做法：改为"先立规矩，再开工"。① 用一次生成产出角色参考表（案例 2-A 的成果）；② 按表 2-18 把 6 帧逐一写成画面任务表，特别是写死"判定标准"；③ 冻结基线参数（同一平台同一模型档位、16:9、固定的引导强度、每帧 4 张）；④ 一晚跑完全部 6 帧，共 24 张；⑤ 用"质检四看"筛掉 3 张明显不合格的（一张指节异常、一张有乱码文字、一张与 K1 构图撞车），对其中 2 张做局部重绘、1 张改图生图微调；⑥ 6 张定稿连同提示词、参数、种子一起归档。

结果：六张图第一次成为一个"组"——色调一致、人物是同一个人。最让她意外的是第 ④ 步：**批量生成反而比"一张一张精雕"更快**，因为筛选发生在生成之后，比较发生在同一批的同一次分布里，直觉判断也更准。

复盘：①"先定角色，再定画面"的顺序不能反——小月第一版绕过了参考表，结果 K2 和 K5 是两个人；②基线参数的意义是把"风格漂移"这项风险从流程里拿掉；③参考表里的人物是正面站姿，而 K2 需要侧身——参考图解决"长得像"，姿态交给提示词或姿态控制，两者分工不要混淆；④K5 的人影太小，后来干脆去掉了五官细节，只保留轮廓——这是"以退为进"的取舍，值得写进复盘。

◆ 外部案例：2025–2026 年，工业界在做什么

课堂作业和真实产业的距离，比你想的要近。下面三条事实放在一起看，能拼出一张相当清晰的产业图景：

- **它已经在生产线上。**2026 年超级碗出现了首个 AI 生成广告（ComfyUI 生态案例）；"ComfyUI artist" 甚至已经成为一个招聘职位名。同年 8 月 17 日，**美国电影协会（MPA）与字节跳动就 Seedream、Seedance 等模型的知识产权保护达成合作并签署谅解备忘录**——版权方与生成模型方开始坐下来谈规则了。
- **它摔的跟头和你摔的是同一个。**2025 年，荷兰麦当劳的 AI 圣诞广告因**人物肢体扭曲**被下架。注意这句：不是"风格不对"，也不是"不好看"，而是**肢体扭曲**——和你在作业里遇到的"六指""手与手臂接不上"是同一类失败。动态场景（第 3 章的内容）放大了这个问题：静止的一帧里能糊过去的错误，动起来就藏不住了。
- **合规正在被写进产业合作。**MPA 与字节的谅解备忘录、日本 17 家出版机构针对 Sora 2 的联合声明（2.15 节详述），都说明"素材从哪来、生成物怎么标"已经不是一个可选项。

这三条事实给课堂的启示是：你们在课上练的"画面任务表、质检四看、归档参数"，正是产业界交付流程的缩小版。区别只在于规模——而方法的形状是一样的。

2.13 失败案例诊所：症状、成因与处方

生成式图像创作有一个残酷的真相：你花在"修"上的时间，往往会超过花在"生"上的时间。因此，一个成熟的创作者与一个新手最大的区别，不在于会不会写提示词，而在于**能不能一眼看出问题出在哪一层**——是数据层的先天不足？是参数的锅？还是你在要求一件模型做不到的事？

本节开出一间"诊所"：先给一张失败模式的解剖图与一张总表，再对五个最常见的失败模式逐一给出症状、成因与处方。

图 2-17　失败模式解剖。任何一次翻车都可以沿"症状 → 四层成因 → 三层处方"走一遍。先判断问题在哪一层，再决定是修、是重掷、还是换模型——这个顺序能省掉大半无效操作。

表 2-19　失败模式与成因总表（"失败诊所"用表）

症状	成因（分层）	处方
多余手指 / 六指	数据层：手在训练图里面积小、模糊、常被遮挡；统计层：学的是"4–6 根类似线条"而非"五根"的硬规则，遮挡样本把物体边缘误学为手指；生成机制：局部拼接、无回溯修正，错了只会被强化；噪声：手掌边缘被误判为手指起点	负向提示词只能缓解；**局部重绘必须改变模型可见信息**（蒙版完整覆盖问题手指 + 提示词写清数量与朝向）；注入姿态约束；长期方案是换对手部学习更充分的新模型；**高清化会放大问题**
肢体 / 关节异常	姿态空间巨大（单手就有 21 个关键点）＋强自遮挡＋模型没有解剖学硬约束	**重新生成**而非局部重绘；降低手部动作难度（自然下垂、插兜优于比心、握笔、捧杯）
手与物体粘连	边界信息完全缺失，手与杯子/头发/袖口糊成一团	顺序铁律：**先清场，再重画**——把粘连团连同所粘之物一起消除，得到干净画布再重绘；预防性做法是先生成空手图再单独合成物体
手部透视错误	手与手臂无法连接、掌心方向与手臂相悖——**全局关系错误，局部窗口装不下**	重新生成；或用"换手法"：从同批生成中取一只正确的手，抠图、对齐透视与光照方向、贴回，再用 AI 修图收缝
单个别扭的指关节	数量与透视都对，只是关节形态不自然	用**液化工具**推一下（几秒钟、不引入新随机性），而不是局部重绘——每次重绘都是全新采样，可能"修好一处又坏了另一处"
文字崩坏 / 乱码	早期模型缺少字形级监督；分辨率不足；复杂排版与长文本的逻辑一致性差	换中文优化模型（Qwen-Image 系、Seedream、GLM-Image、Kolors）；提高分辨率；减少单图文字密度；大段文字交后期排版兜底；定稿前**逐字核对**
光影不一致 / 过曝	**CFG 过高**导致饱和度与微对比急升、高光溢出、阴影过重；扩图/生成填充引入的新区域与原图光照不匹配	降低 CFG（负向提示词只能缓冲，不能替代）；CFG-rescale（φ=0.5–0.7）、APG、dynamic thresholding，或把引导限制在中间区间；扩图后逐项检查边缘、光影
风格漂移	风格描述过粗时模型"按自己的文艺感填空"；跨模型/跨版本切换；参考图带杂物；模型大版本更新让风格词稳定性重新洗牌	用标准化程度更高的英文风格词；冻结基线参数；清理参考图；版本更新后**重新拉基线**
双胞胎脸 / 角色混乱	多主体时模型按"平均脸"采样；提示词没给每个角色独立的可区分特征	用参考图/主体参考为每个角色锁定身份；多参考图分工；为每个角色写独立特征
低 CFG 的漂移与数量错误	CFG 过低时模型优先依赖先验而非提示词	适度提高 CFG，或改用结构化控制（参考图/姿态/边缘）
非方形画幅的元素重复	扩散模型训练分辨率较单一，生成 16:9 等非方形画幅时容易重复元素	优先选模型**预设支持**的宽高比；或使用对非方形优化过的模型
"一眼假"	服饰上的乱码文字、手部/动作异常、人脸比例不符	这三类特征也是当下最实用的**鉴定线索**——把它当课堂"找茬"练习，也把它当自己的发布前自查表

◆ 五个真实病例

🩺 失败诊所 F2-1 "六根手指"与它的四次重绘

症状：小月生成的 K2 里，诗人垂在身侧的手有六根手指，其中两根从同一指根分叉出来。她挂了蒙版重绘——第一次修好了分叉，又冒出一根；第二次数量对了，指节却像被拧过；第三次整只手糊掉了。

成因：回到图 2-17 的四层链路看：模型学到的不是"手必须有五根手指"的硬规则，而是"手的图像通常包含 4–6 根类似线条的结构"；而局部重绘在局部同样是**重新加噪再独立采样**——所以"修好一处、坏了另一处"不是运气差，是机制使然。此外她犯了一个经典错误：把局部重绘当成"纠正数量"的工具，而没有改变模型可见的信息。

处方：①重绘时让蒙版**完整覆盖问题手指与粘连边界**，并把"数量与朝向"写进局部提示词（"五根手指清晰可数，拇指与其余四指分开，手背朝向镜头"）；②两次重绘仍不对，**立刻停止重绘**——这是一条止损纪律；③改用"重新生成"或"换手法"；④预防层面：把动作改成"自然下垂"甚至"插兜"（不露手指就没有计数问题，社区里把它戏称为小作弊）；⑤长期看，换一个对手部学习更充分的模型，比任何技巧都有效。**记住：高清化只会让这只手更清楚地告诉你它有六根手指。**

🩺 失败诊所 F2-2 海报上"被吃掉"的两个字

症状：小舟做课程海报，想让模型直接把标题"春江花月夜"写进画面。第一版把"夜"字渲染成了不存在的字；第二版五个字全对，但角落的活动说明小字乱码成一片；第三版她把 12 条活动信息全塞进图里——结果连最好用的模型也没顶住：一张"AI 模型入门指南"风格的网格卡片里出现了 5 处文字错误。

成因：文字崩坏属于"分辨率 + 监督 + 排版复杂度"的复合问题：早期模型缺少字形级监督；分辨率不足时小字必然糊；而**高密度信息图与长文本的逻辑一致性**，即使用 2026 年最好的模型也仍是主要失分点——这不是她的使用方法错了，是她给模型派了一个尚未成熟的活。

处方：①换中文优化模型（Qwen-Image 系、Seedream、GLM-Image、Kolors 等，中文汉字渲染是它们的专门优势）；②提高分辨率，并**减少单图文字密度**——标题类大字可以生成，成段小字不要交给模型；③大段文字一律后期排版兜底；④定稿前把图放大到 1:1 逐字核对。**一句话的操作标准：图里的字要"少而准"，多而不准等于没写。**

🩺 失败诊所 F2-3 "过曝的月亮"：一个参数引发的全组返工

症状：小月为了"更贴提示词"，把引导强度从甜点区拉到 14。生成的月亮确实更圆更亮，但整张图饱和度与微对比急升、高光溢出成一团白斑、人物皮肤出现塑料感、江面边缘生硬——六张里有三张因此报废。

成因：典型的"CFG 过高"事故。引导强度不是质量旋钮而是**权衡旋钮**：它把"有条件"与"无条件"两次预测的差值放大，放大提示词影响的同时，也放大了这些副作用；到 12 以上结构开始崩坏，21 以上通常直接崩。她试图用负向提示词写"不要过饱和"来补救——事实是负向提示只能缓冲，**不能替代降低 CFG**。

处方：①把引导强度退回模型甜点区（对照表 2-4 的参考区间，但以你自己实测为准）；②若需要更精细的控制，使用 CFG-rescale（φ=0.5–0.7）、APG、dynamic thresholding，或把引导限制在中间采样区间（如 `guidance_interval=[0.2,0.8]`）；③扩图或用生成式填充后，必须逐项检查新区域的边缘、光影与原图是否接得上；④把"每次只改一个参数"写进实验纪律——她这次是全组六张一起改的，代价自然也是六倍。

🩺 失败诊所 F2-4 "第三张开始不像一套"

症状：小月第一版六张关键帧，前三张是清冷青蓝，第四张开始偏暖、颗粒变粗，第六张干脆成了另一种画风。她的提示词里风格部分只写了"中国风，唯美"。

成因：三层叠加。第一，**风格描述过粗**——"中国风，唯美"这样的词留白太大，模型会"按自己的文艺感填空"，而每次填空的答案不同；第二，她中途换过一次工具做对照，跨模型切换会让风格词**重新洗牌**；第三，参考图里带了原图背景的杂物，模型把杂物也当成了风格线索。此外还有一条客观规律：**模型的一次大版本更新，会让所有风格词的稳定性重新拉基线**。

处方：①用标准化程度更高、颗粒度更细的英文风格词（详见 2.10 节中英对照）；②冻结基线参数，对照实验单独开一组、不混进正式批次；③参考图清理杂物，必要时先抠图再挂；④把"版本变更后重拉基线"写成项目规范。这条处方背后还有一条产业级的思路值得记住：Seedream 4.0 用**统一架构承载生成与编辑**，官方给出的理由之一正是"避免多模型切换导致的风格漂移"——**工具选型本身就是在预防风格问题**。

🩺 **失败诊所 F2-5** 两张"平均脸"：多人同框的身份事故

症状：小组想做一张"江边对饮"的合影：诗人与友人对坐。生成的结果里，两个人像同一个人的不同发型；换了三版，每一版都是两张"平均脸"。

成因：多主体时模型倾向于对每个角色采样一张"平均脸"，而提示词里除了"友人穿深色衣服"之外，没有任何可区分的身份信息——模型无从区分"谁是谁"，自然长成了双胞胎。这个现象的现实版本相当有名：2026 年呼和浩特摄影赛一件 AI 作品被识破的证据之一，就是"图中的环卫工**面容高度相似**"。

处方：①为每个角色准备各自的参考图或主体参考（多参考图分工），把身份锁定从"描述"升级为"给定"；②为每个角色写独立可区分特征（脸型、发式、关键配饰），而不是只写服装颜色；③若工具支持多角色参考，一图一角色、编号对应；④出图后用"四看"里的"看脸"逐人核对。多人场景的一致性，本质上是"给每个角色提供独立身份来源"的工程问题。

◆ 两条止损纪律：什么时候承认"就这样吧"

失败诊所的最后，是两条比任何技巧都重要的纪律——它们是第 1 章"一次只改一个变量"在图像领域的延伸：

1. **重掷不会因为"再试几次"而变容易。**不改动模型可见信息的重掷，本质上就是独立采样——五次失败不代表第六次更接近成功。所以**止损线应当按成本来设，而不是按胜率来设**：先估算"手工修复这张图需要多久"，再和"已经花在重掷上的时间"比较；当两者接近时，继续重掷就是负收益。这条纪律的价值在于，它把"不甘心"这个情绪变量，换成了一个可以做算术的决策变量。
2. **每轮只改一个变量，且通过标准写在开始之前。**提示词、手部动作、构图、参考图——一轮只动一个，否则无法归因。做 A/B 实验时，通过标准要在开始前就写死（例如在 1:1 像素视图下：手指可数、指节数量正常、掌心朝向与手臂一致），每组跑两遍、中间留出间隔，以判断差异是"变量造成的"还是采样噪声造成的。

还有一句必须坦率的话，请所有同学记在实验记录的第一页：本节与表 2-19 给出的所有"处方"，都是方法，不是数字承诺。结论会随模型版本更新而失效——这就是本书不提供任何"通过率表格"的理由。你要带走的是**判断问题的框架**，而不是一张会过期的成绩单。

2.14 动手实践：从文生图到关键帧的四个阶段

本节把全章内容落到手上。练习按四阶段递进：**文生图 → 图生图 → 局部重绘 → 条件控制**，中间插入两组**参数对照实验**（一次只改一个变量：引导强度、采样步数、种子），最后以主线任务收束。所有练习都要求同一个动作：记录——提示词、参数、模型版本、日期、结论。

◆ 🛠 实践一：文生图起步——同一提示词的三次采样（30 分钟 • 课堂）

基础

选一个国内网页工具并**固定下来**（即梦、豆包、通义万相等，选能登录、能保存历史记录的那个）。用一条四要素提示词描述"月色下的江面"，参数全部保持默认，连续生成 **3 次**。

观察并回答：三次结果像不像？哪里像、哪里不像？"月亮的位置""水面的反光""雾气的形状"分别变了吗？

提交物：3 张图 + 一段 80 字以内的观察记录。这段记录的结论应该是：**同提示词、同参数下，每次生成仍是一次独立采样——差异是机制的一部分，不是操作失误。**

进阶

给这条提示词加一个约束（例如"画面留白占三分之二"），再生成 3 次。观察：约束是否每次都生效？约束越多，三次结果的差异是变大了还是变小了？把答案写进记录——它是第 2.7 节"提示词越具体、约束越强，但冲突指令会互相抵消"的现场版本。

🛠 实践二：参数对照实验 A——引导强度（40 分钟 • 课堂）

进阶

固定同一个提示词（建议用《春江花月夜》画面），**其他一切不变，只改引导强度**，按下表跑三组，每组生成 2 张（间隔几分钟再跑第二张，以判断噪声）。

表 2-20　参数对照实验设计（一次只改一个变量）

实验	变量（只改这一个）	三组取值	固定不变项	观察点	通过标准（开始前写死）
实验 A	引导强度（CFG）	偏低 ≈3 ｜ 甜点区（先按 4–7 试一个值） ｜ 偏高 ≈12	提示词、模型档位、分辨率、采样步数、宽高比、每组的采样次数	提示词遵守度、饱和度、高光是否溢出、边缘是否生硬、皮肤质感	能同时说清"低档丢掉了什么"与"高档多出了什么副作用"；能指出哪一档最接近"提示词遵守度高、色彩光照不戏剧化"
实验 B	采样步数	少步档 ｜ 平台默认 ｜ 约 2× 默认	同上（引导强度固定在实验 A 选出的值）	细节、收敛程度、耗时、成本；若用蒸馏/少步模型，观察"超出蒸馏轨迹"后是否反而变差	能给出"边际收益从哪一步开始迅速递减"的实测感受；能复述"步数要与该模型的蒸馏设定匹配"
实验 C	种子	固定种子 ｜ 换种子 ｜ 固定种子 + 只改提示词中的一个词	其余全部不动	哪些元素被锁住了、哪些仍在变；"固定种子"到底保证到什么程度	能实物演示三条边界：换模型/分辨率/采样器后同一颗种子不再复现；改蒙版会改变输出，即使种子固定

挑战

故意把引导强度拉到 **≥12** 生成一张，把它作为**反例存档**，在图注里写清楚崩坏的方式（结构崩、色带、高光溢出……）。反例存档的价值在于：以后见到类似症状，你能立刻联想到"是不是引导强度过高"。

🛠 实践三：图生图与重绘幅度（30 分钟 • 课堂）

进阶

取实践一里你最满意的一张图，做三组图生图，**只改重绘幅度**：**0.25 / 0.5 / 0.8** 各一张。

逐张回答：主体身份还在吗？构图还在吗？画风变了多少？把三张按"改动强度"排序，写一行结论：**重绘幅度就是"加了多少噪声"，幅度越大，原图残留信息越少、模型自由发挥越多。**

🛠 实践四：局部重绘——给蒙版立规矩（40 分钟 • 课堂）

1. **制造一个问题。**用实践一的图，或直接生成一张"人物手持杯子"的画面（注意：这个动作本身容易出问题，正好当教具）。
2. **遵守蒙版三条规则重绘。**①蒙版比物体边缘略大一圈软边；②把物体造成的后果（阴影、反光）也框进去；③复杂主体分两趟——先修物体、再修阴影。
3. **故意违规一次。**把蒙版画得紧贴物体再重绘一次，观察"外溢的边"是怎么出现的——对比出真知。
4. **验证种子的边界。**固定种子，只改动蒙版形状，再生成一次，记录输出是否变化。（预期结论：改蒙版会改变输出，即使种子固定。）

挑战

如果你用的是**纯提示词式**编辑工具（不接受蒙版文件），请完成同一任务并总结：纯提示词式修不好的那一类问题是什么？（提示：需要"重画边界"的那一类。）把结论写进记录，并说明为什么"蒙版式与纯提示词式的失败方式不同"。

🛠 实践五：条件控制——让"必须像什么"说话（30 分钟 • 课堂 + 课后）

进阶

找一张构图清楚的参考图（风景或人物均可），用工具内的**边缘控制**或**姿态控制**复现其构图，生成 2 张，然后**只改控制权重**（例如 0.4 与 0.8）再来一轮。

回答三个问题：①权重 0.8 时，新图与参考图"像"在哪里？②权重 0.4 时，模型自由发挥在哪些部分？③"权重越大越好"这句话错在哪？

挑战

组合使用两种控制（参考图锁角色 + 姿态锁动作），生成一张"指定的动作 + 指定的人"。如果两次都失败，按 2.13 节的止损纪律记录：你花了多少时间、手工修复的预估耗时是多少、你决定继续还是止损——**把决策过程本身写成记录。**

🛠 实践六：主线推进——角色设定图与关键帧（60 分钟 • 课堂 + 课后）

1. **定角色。**按案例 2-A 的方法产出一张"角色设定图"（正面、全身或头肩、均匀布光、中性灰底），不满意就迭代，**定稿后不再改动。**
2. **写任务表。**按表 2-18 的格式，为 4–6 个镜头各写一行：画面意象、景别、生成要点、一致性与控制、主要风险——**判定标准必须在生成前写死。**
3. **冻结基线参数。**同一平台、同一模型档位、同一分辨率（16:9）、固定引导强度、每帧 4 张。写进记录表。
4. **批量生成与筛选。**一晚跑完全部镜头，用"质检四看"（手 / 字 / 光影 / 脸）筛出定稿，问题图按"局部重绘 → 图生图 → 重新生成"的代价顺序处理。
5. **归档。**提示词、参数、种子、模型版本、日期、一句复盘，缺一不可——第 1 章的"创作档案"在这里迎来第一次大考。这些素材将在第 3 章作为视频生成的首帧直接复用。

所有实验共用一个记录表模板。请把它复制到你的创作档案里，**每个实验一行，并强制填写"模型版本 + 日期"**——原因已经说过：结论会随版本失效，没有版本号的结论等于没有结论。

表 2-21　参数对照实验记录表（模板，复制到创作档案）

实验编号	变量（只改这一个）	取值	固定不变项	观察点	结论（附模型版本 + 日期）
例：E-01	引导强度	3 / 5 / 12	提示词 · 模型 · 分辨率 · 步数	遵守度 / 饱和度 / 高光 / 边缘	例：本模型甜点区在 5 附近；12 时高光溢出、结构开始崩（模型 vX，2026-10-09）
E-02	—	–	—	—	填写你的结论
E-03	—	–	—	—	填写你的结论
E-04	重绘幅度	0.25 / 0.5 / 0.8	原图 · 提示词 · 模型	身份保持 / 构图保持 / 画风变化	填写你的结论
E-05	控制权重	0.4 / 0.8	参考图 · 提示词 · 模型	结构相似度 / 风格自由度	填写你的结论

2.15 课程思政：用图像讲中国故事，也守好取材的边界

本章的思政主题有两半，一半是**创作的抱负**——中国的创作者应该用这门新技术讲什么；另一半是**创作的底线**——素材从哪来、作品怎么标、别人的东西能不能用。两半合成一句话就是：技术可以全球采购，母题必须自己说话；自由不是无边界的，边界正是专业性的组成部分。

◆ 为什么《春江花月夜》是一个好选择

先回到贯穿案例。小组选择用张若虚的《春江花月夜》做两分钟短片——这不是"因为古诗简单"，恰恰相反，它是"用最前沿的工具，做最中国的表达"的一次练习，而且它有两层现实红利。

第一层是**技术红利**。你可能已经注意到 2.10 节里的一条规律：**文化语境用母语最稳**——春节场景用中文提示词最好（红梅、窗花、金箔保持完整），中秋场景用中文能保持月饼与茶席的构图。原因很简单：模型在训练数据里见过大量真实的中文文化场景，当你的母题与训练数据里的真实存在重合时，"创造"就变成了"唤起"，难度大幅下降。中国古典意象（江、月、花林、孤舟、楼台）正是中文优化模型的优势区。

第二层是**责任红利**——这个词是本书的私货，但值得解释。当全世界的创作者都在用几秒钟生成"一个差不多的古风画面"时，**真正的稀缺品不是画面，而是准确的画面**：楼的形制对不对？袖口的形制是哪个朝代的？"秉烛游"该用烛还是灯？这些细节无法靠"泛东方审美"蒙混，只能靠查证。谁愿意查证，谁的作品就有别人模仿不了的可信度。**技术让"生成"变得廉价，恰恰让"考据"变得昂贵。**

反过来，不做考据的代价也实实在在地摆在眼前。本书编写时检索到的两条事实值得所有同学记住：其一，模型对文化符号存在"按训练数据里的刻板印象填空"的倾向——中文"复古"指向 1980–90 年代港风，翻译成英文后"vintage"却指向 1950–70 年代美式；春节意象翻译成英文会漂向"泛亚洲节庆"，中秋意象会漂向日本茶道。其二，这种漂移不是抽象的理论：2025 年"岭南杯"全国古树名木摄影大赛的核查中，一件标注为海南的获奖作品，实际上是**印度尼西亚松巴岛**的树木图片——"用错地方"在专业赛事里是会被撤销奖项的硬伤。

> 🇨🇳 **课程思政 · 用图像讲好中国故事：三层操作**
>
> **第一层：给意象配"考据"。**中国古典意象（月、江、花、舟、楼台、衣冠）在提示词里只是一个词，在成片里却是一整套形制。做法是把"考据"变成流程的一步：每张关键帧旁边附一行"形制说明"——画的是哪个时代的楼、什么形制的衣冠、依据哪些图片资料。本节主线案例的 K4"楼台"就是这么做的：小舟花了一个小时查唐宋楼阁的图片资料，再照着写提示词。用参考资料与参考图锁定形制，是防止"泛东方拼贴"最有效的手段。
>
> **第二层：不越"I 化的熟悉感"。**有些画面"看起来很中国"，其实是别的文化：灯笼堆叠、红金配色、飞檐乱贴——这不是中国，这是训练数据对"东方"的平均印象。判断标准很简单：能不能说出每个元素"是什么、出自哪、为什么在这里"。说不出的，删掉。**用刻板印象省下的时间，最后都会在观众的质疑里还回去。**
>
> **第三层：请"懂的人"审看。**涉及文化细节的作品，定稿前请一位相关专业的同学或老师过目：形制、仪轨、称呼、器物。这一条听起来像常识，但它和"写完代码请人 review"是同一种职业习惯——评审不是不自信，是专业流程的一部分。

◆ 合法取材：四个判例划出的边界

接下来是本章思政的另一半，也是数字媒体创作者最容易踩坑的地方：素材与生成物的**权利边界**。中国的法院已经用一组判决把边界画得相当清楚——下面这四个案子，建议逐条读完，因为它们直接决定了你的"创作档案"该怎么建。

表 2-22　AI 生成图像相关的四个中国判例（检索口径：2026 年 10 月）

案件	法院与时间	要点	对作业的启示
李某诉刘某案（"春风案"，媒体称 AI 文生图著作权第一案）	北京互联网法院 2023 年 11 月一审，2024 年 1 月报道已生效（双方均未上诉）	李某用 Stable Diffusion，通过输入提示词、设置**迭代步数、图片高度、提示词引导系数、随机种子**等参数生成人像图并发布；刘某在配图中使用并**截去署名水印**。法院认定李某"从构思到最终选定图片进行了智力投入"，涉案图片具备"智力成果"与"独创性"要件、**属于美术作品**；李某是作者；**著作权法规定的作者限于自然人、法人或非法人组织，AI 模型本身不能成为作者**。判赔经济损失 500 元并赔礼道歉，同时强调**需个案判断**	**保留参数记录 = 保留作者身份的证据**。你写进创作档案里的步数、引导强度、种子，正是本案中法院采信的关键事实
周某诉某科技公司案	北京互联网法院	周某主张其 AI 生成的"某猫"图片被某公司用于推广，法院**驳回**其主张；重要理由之一是"**周某并未提交涉案图片在 AI 软件中的创作过程记录**"，证据无法证明独创性，涉案图片不构成"作品"	与李某案的**分野就在证据**：有没有创作过程记录、能否体现个性化表达与智力投入。这就是本书从第 1 章起强制要求"创作档案"的法律原因
奥特曼案	广州互联网法院 2024 年 2 月判决	被描述为"**中国法院作出全球首例生成式 AI 服务侵犯著作权的生效判决**"。被告 AIGC 平台可由简单关键词生成奥特曼图片；法院认定构成复制权与改编权侵权、平台未尽合理注意义务；判赔金额不高，判决指出行业"正处于发展的初期，需要同时兼顾权利保障和产业发展"	**不要生成现有的知名 IP 形象**（动漫角色、品牌形象、明星面容）。"输入几个关键词就能出图"的便利，不改变形象本身的权利归属
北京通州法院案	2025 年 6 月 13 日判决	涉及利用 AI 技术侵犯著作权；与奥特曼案一样触及"AI 辅助二创是否产生新作品"的问题	二创不是免罪牌：在他人作品之上"再创作"，仍要先过权利关

把四个案子压成一句操作口令：输入要干净，中间要留痕，输出要标识。"输入要干净"指不喂他人受保护的作品当作风格复制材料、不生成现有 IP 形象；"中间要留痕"指创作过程记录（提示词、参数、过程截图、日期）；"输出要标识"指 AI 生成物的显式与隐式标识。第三条不是可选项，它是两套正在强制执行的法律体系的要求。

关于"风格模仿"，还要补一条 2025 年的产业风波。当年 OpenAI 的图像生成能力可以"只需一句话和几秒钟就生成宫崎骏动画"，引发吉卜力工作室等日本 IP 出版公司抗议；吉卜力等公司通过日本内容产品海外流通促进机构（CODA）向 OpenAI 致函，要求停止未经授权使用相关内容训练 AI 模型，并指其 Sora 2 已生成大量与日本现有内容产品相同或高度相似的影像。日本方面随后形成合力：包含小学馆在内的 **17 家出版机构**以及日本动画协会、日本漫画家协会于 10 月 31 日联合发表声明；10 月 10 日，时任日本知识产权担当大臣城内实表示日本政府已要求 OpenAI 不得实施侵犯著作权的行为。OpenAI 提出的"选择退出"（权利人须主动声明"不得使用"）机制，被 CODA 明确反对——它与日本著作权制度"使用著作物原则上须事前许可"存在根本冲突。**这场风波的核心问题正是："模仿一种画风"算不算越界？目前仍在全球范围内争论，尚无统一答案。**对课堂的启示不是站队，而是**谨慎**：把"模仿某位在世艺术家/某个工作室的独有风格"从作业要求里划掉，改成"概括其特征的通用风格词"（例如"水彩质感、柔和边缘、低饱和"），既安全，也更专业。

◆ 发布前的合规检查：把法规变成操作

法规条文读起来遥远，但落到操作上只有几条。先看现行要求：中国的《人工智能生成合成内容标识办法》（网信办、工信部、公安部、广电总局联合发布）**自 2025 年 9 月 1 日起施行**，配套的强制性国家标准 **GB 45438—2025 同日实施**；欧盟《人工智能法案》第 50 条的透明度义务则于 **2026 年 8 月 2 日**正式生效并可在全欧盟/EEA 执行。两边的要求用一张表对照如下。

表 2-23　发布前合规检查清单（图像作品）

检查项	具体要求（要点）	你的操作
① 显式标识（中国）	图片等生成合成内容，服务提供者应添加显式标识：在图片适当位置以文字/图形呈现、可被用户明显感知；提供下载、复制、导出时须确保文件含满足要求的显式标识。国标规定：**采用文字提示、位于图片的边或角、字型清晰可辨、文字高度不应低于画面最短边长度的 5%**，并应同时包含"人工智能"或"AI"要素以及"生成"和/或"合成"要素（如"AI 生成"）	课程对外发布的作业一律执行：导出后确认标识存在，且在缩略图尺寸下仍可辨认
② 隐式标识（中国）	在文件**元数据**中添加生成合成内容属性信息、服务提供者名称或编码、内容编号等；**数字水印被列为"鼓励"而非强制**	不要清理工具写入的元数据；转发、压缩、截图前先在档案里留一份原始文件
③ 不得恶意处理标识	任何组织和个人**不得恶意删除、篡改、伪造、隐匿标识**；传播平台对未标识但检测到 AI 痕迹的内容需添加风险提示	把"保留标识"写进自己的发布流程——李某案中被截去的正是署名水印，那已构成侵权
④ 发布场景自查（国际对照）	欧盟第 50 条：提供者须确保合成图像输出以**机器可读格式**标记并保持可检测，标识应有效、可靠、稳健、可互操作，技术措施应在设计和开发阶段内嵌；部署者发布**深度伪造**，或可能被误认为真实、涉及真实人物/地点/事件的 AI 生成图像，必须**以清晰可见方式披露 AI 来源**；艺术、创作或讽刺作品可豁免深度伪造标识；C2PA 规范被作为过渡参考。**罚则最高 1500 万欧元或全球年营业额 3%（取高者）**；**2026 年 12 月 2 日**为已上市生成式 AI 系统机器可读标识的过渡期截止	涉及真实人物/地点/事件的图，无论发在哪，都加清晰可见的 AI 披露；不要生成"看起来像真实新闻照片"的画面
⑤ 素材来源	判例层面：李某案采信"智力投入"，周某案因缺少创作过程记录败诉；奥特曼案确认生成他人 IP 形象构成侵权	每个素材一行来源记录（URL、授权方式、日期）；生成记录（提示词、参数、种子、模型版本、日期）随成品归档
⑥ 真人权利	涉及真人肖像、面容的生成与合成是高风险区	不生成可识别的真实人物（除非本人授权且用途明确）；不使用同学生成他人"图生图"肖像
⑦ 文化准确性	文化符号误用与刻板印象既是伦理问题，也是作品质量问题	按"三层操作"执行：形制说明、拒绝泛东方拼贴、请懂的人审看

最后，把这套检查放在一个真实的场景里检验。2025–2026 年中国多起摄影赛事撤销 AI 作品奖项的事件（呼和浩特全民摄影双月赛《洒乐园林》2026 年 7 月 14 日被通报"确系 AI 生成"、取消资格并暂停赛事整改；《大众摄影》"五腿牛"照片 2026 年 7 月 17 日撤下，杂志随后修改征稿规则，要求 AI 生成影像**主动声明并明显展示 AI 标识**，且不接受纯粹文生图影像；"岭南杯"撤销二等奖并永久取消参赛资格；南丰摄影赛取消获奖资格且空缺奖项不再递补）——这些事件的共同点不是"AI 被讨厌"，而是**隐瞒**被讨厌：赛事处罚的是"不声明、伪造拍摄信息、伪造地点"，而不是"使用了 AI"。把 AI 标识亮出来，你不仅不违规，还比多数人更专业。这也正是本书要求"作品内可见标识 + 档案留痕"双轨并存的原因。

◆ 把这一章压缩成七句话

1. 从 GAN 到扩散：GAN 靠"生成器对抗判别器"博弈，训练不稳、易模式崩溃；扩散把"生成"拆成"逐步去噪"——正向过程按固定规则加噪、反向过程由网络一步步减噪，**训练 = 学会去噪，生成 = 从噪声逐步雕刻。**
2. 潜空间：扩散不直接在像素上工作，而在 VAE 压缩后的潜空间里工作；8× 空间压缩带来约 64 倍的算力节省，代价是有损压缩会留下痕迹。
3. 架构演进：U-Net 的跳连 → DiT 把潜变量当成视觉 token（AdaLN 注入条件）→ MM-DiT 的文本/图像双流双向注意力；2026 年又出现单流、混合堆叠、MoE 等分支——架构之争并未结束。
4. 文本条件与 CFG：引导强度是**权衡旋钮，不是质量旋钮**；负向提示词只是"无条件分支"的替代品，能缓冲、不能替代降低 CFG；甜点区随模型不同（记得查官方推荐或自行实测）。
5. 流匹配 / 整流流：把"噪声到图像"的路径拉直，2024 年后成为主流；但要说清"框架"与"实现"两个层次，"扩散被淘汰"的说法并不准确。
6. 参数与编辑：一次只改一个变量；图生图用重绘幅度控制"改多少"，局部重绘用蒙版获得"重画边界"的机会，条件控制用结构图约束"必须像什么"，角色一致性走参考图/LoRA/提示词块三条路。
7. 失败与合规：失败可沿"症状 → 四层成因 → 三层处方"归因，止损线按成本设而不是按胜率设；技术可以全球采购，母题必须自己说话——**输入要干净、中间要留痕、输出要标识**，创作档案既是经验的沉淀，也是法律的证据。

◆ 自测

❓ 自测 1：用"训练 = 学会去噪，生成 = 从噪声逐步雕刻"解释扩散模型的原理。为什么说"生成不是一次画出来的"？

答案：训练阶段做两件事：把一张训练图按固定规则逐步加噪（正向过程），让网络反复做"看这张带噪的图、报出里面加了什么噪声"的练习；练到最后网络成为一个"噪声侦探"。生成阶段则把这个本领反着用：从一张纯噪声图开始，让网络预测其中的噪声、减掉一部分，再预测、再减——重复几十步，噪声一点点褪去，图像一点点显形，就像从石头里"雕刻"出来。之所以说"不是一次画出来的"，是因为每一步的输入都是上一步的输出，模型并不持有整幅画的全局蓝图，它每一步只负责"这一步该少掉哪一部分噪声"。这个机制解释了本章的许多现象：为什么步数是"循环次数"而不是"精度档位"（2.7 节）、为什么局部重绘会"修好一处坏了另一处"（2.13 节 F2-1）、为什么少步蒸馏模型跑超出设定的步数反而变差（超出蒸馏轨迹）。

❓ 自测 2：扩散为什么要在潜空间里做？"8× 压缩带来约 64 倍算力节省"是怎么来的？有损压缩会带来什么后果？

答案：在像素上直接做扩散太贵：一张 1024×1024 的 RGB 图约有 314 万个数字，每一步去噪都要让网络处理这么多数字，还要重复几十步。潜空间的做法是先训练一个 VAE：编码器把图像压成小得多的"潜变量"（空间上约 8× 压缩），扩散在潜空间里进行。8× 是**长度和宽度各压 8 倍**，所以面积（也就是要处理的数值量）压低约 8×8 = 64 倍——这就是"约 64 倍算力节省"的来源。后果有三条：①解码器要负责把潜变量还原成像素，训练得不理想时表现为细节发糊或出现伪影；②压缩是**有损**的，微小文字、细纹理这类"高频率信息"最容易被压掉，这客观上加重了文字渲染与手指等比细节的难度。③"表示空间"本身还在演化（不同模型的 VAE 通道数、压缩倍率不同），并直接影响一个模型擅长什么——这也解释了为什么"换模型"有时比"调参数"更能解决某类问题。

❓ 自测 3：小月为了"更贴提示词"，把引导强度从 5 调到 14。请预测会发生什么；如果她再写一句负向提示词"不要过饱和"，会有帮助吗？为什么？

答案：CFG 把"有条件预测"与"无条件预测"的差值放大：$\varepsilon_{\text{引导后}} = \varepsilon_{\text{无条件}} + w \cdot (\varepsilon_{\text{有条件}} - \varepsilon_{\text{无条件}})$。w 调到 14 后，提示词遵守度会更"硬"（细节更贴、线条更锐），但同时副作用全面放大：饱和度与微对比急剧上升、高光溢出、皮肤出现塑料感、边缘生硬；到 12 以上结构开始崩坏，21 以上通常直接崩——她的图很可能出现"过曝的月亮 + 生硬的江面"。负向提示词"不要过饱和"只能算缓冲，**不能替代降低 CFG**：它的机制是把"无条件分支"替换成"对负面概念的预测"，只是间接地往外推一点，而副作用来自引导本身的放大机制。正确处方：把引导强度退回甜点区（不同模型的甜点区不同，参考表 2-4 并自行实测），需要更精细控制时使用 CFG-rescale、APG、dynamic thresholding 或把引导限制在中间区间。

❓ 自测 4：小月给一张"六指"的图做局部重绘，两次都是"修好这只手、另一处又坏了"。这是运气问题吗？她应该怎么办？

答案：不是运气，是机制。局部重绘在自己的窗口里同样是"重新加噪 + 独立采样"——每次执行都是一次新的随机过程，所以"修甲坏乙"是常态。更关键的是她犯了方向性错误：局部重绘不是"纠正数量"的工具，它的有效前提是**改变模型可见的信息**。正确做法：①蒙版完整覆盖问题手指与指间粘连处（比外缘多留一圈软边），并把物体造成的阴影等"后果"也框进去（蒙版三条规则）；②局部提示词把数量和朝向写清楚（"五根手指清晰可数，拇指与其余四指分开"）；③两次不成即止损——改走"重新生成"、换手法（从同批生成里取一只正确的手），或从源头预防（动作改为自然下垂/插兜、换对手部训练更充分的模型）。还有一条容易忽视的规则：改蒙版会改变输出，即使种子固定——所以"固定种子"不构成"重绘可复现"的保证。

❓ 自测 5：角色一致性有哪三条路径？为什么"使用参考图时，提示词要删掉长相描述"？

答案：三条路径：**A 参考图 / 主体参考**（零训练、约 30 秒级；各家有参考图数量上限，例如 FLUX.2 最多 10 张、Nano Banana 家族最多 14 张）；**B LoRA 训练**（一致性最高，需 15–30 张素材与一次约 1 小时的训练；工程参数如 rank 16–64、学习率 1e-4 起、500–2000 步；角色参考表必须用均匀无阴影棚光 + 中性灰背景，否则会把阴影"烘焙"成下游模型误认的解剖结构）；**C 精确提示词块 / 统一角色文件**（最轻，适合原创角色）。之所以要删掉长相描述：同时给"参考图"和"文字描述"两条身份信息，等于给模型**两个事实来源**——两者不一致时它要"折中"，结果往往两边都不像；头发、眼镜、脸型应当完全交给参考图，提示词只写"在做什么"（情境、动作、镜头）。正确分工记成一句话：参考图解决"长得像"，提示词负责"在做什么"。

❓ 自测 6：李某案与周某案的结果为什么不同？这两起判决对"创作档案"提出了什么要求？发布前还必须补上哪两件事？

答案：两案的**分野在证据**：李某案中，原告能说明自己输入提示词、设置迭代步数、图片高度、提示词引导系数、随机种子并反复调整的过程，法院据此认定"从构思到最终选定图片进行了智力投入"，涉案图片具备独创性、属于美术作品，AI 模型本身不能成为作者但李某是作者；周某案中原告"并未提交涉案图片在 AI 软件中的创作过程记录"，无法证明独创性，图片不构成作品，主张被驳回。对创作档案的要求：**提示词、参数（含种子）、过程截图、模型版本与日期，随成品一起归档**——它平时是经验沉淀，争议时是权利证据。发布前必须补的两件事：①**标识**（显式标识：图片边或角的文字提示，字高不低于画面最短边的 5%，含"AI 生成"要素；隐式标识：保留生成工具写入的元数据，不得恶意删除、篡改、伪造、隐匿）；②**素材与真人核对**（不生成本就存在的 IP 形象与可识别的真人；每个素材有来源行）。把这套动作记成十二个字：输入要干净，中间要留痕，输出要标识。

◆ 拓展阅读

- **ControlNet 原始论文**《Adding Conditional Control to Text-to-Image Diffusion Models》（Lvmin Zhang 等，arXiv 2302.05543，获 ICCV 2023 Marr Prize 最佳论文奖）——读它的第 1 节与消融实验即可，重点理解"为什么要在文本之外再加一条结构通道"。
- **各家官方文档**（优先级高于任何第三方教程）：火山引擎/即梦的模型文档、阿里云百炼的通义万相与 Qwen-Image 文档、Black Forest Labs 的 FLUX 模型卡与许可说明、Midjourney 官方文档页——参数推荐值与商用条款的权威来源，且它告诉你"这一版和上一版哪里变了"。
- **《人工智能生成合成内容标识办法》全文与 GB 45438—2025**——建议在第一次对外发布作品前逐条对照；欧盟《人工智能法案》第 50 条的实施指南（2026 年 7 月 20 日发布）可作为国际对照。
- **北京互联网法院"李某诉刘某案"判决书与相关报道**——通读一遍"智力投入"的认定段落，你会对自己的创作档案产生完全不同的重视程度。
- **《春江花月夜》历代评注与一本中国画论**（如郭熙《林泉高致》）——技法解决"怎么画"，母题解决"画什么"；后者才是作品与作品之间真正的分水岭。

第 3 章

视频生成与编辑

让画面动起来，比让画面好看难一个量级——时空表征、时间一致性、参数与剪辑的完整流水线。

学习目标

1. **知识：**理解视频的时空表征（帧、时间维、数据立方体）与"为什么比图像难一个量级"；能说清潜空间视频扩散、时空注意力、DiT 视频架构三条技术路线的关系，以及时间一致性为什么难（闪烁与漂移的成因）；了解音视频联合生成的进展与"参数语义"背后的工程约束。
2. **能力：**能用文生视频、图生视频、首尾帧、参考图、运镜描述五种条件完成一个镜头的生成；能读懂并调控运动幅度、时长、分辨率、种子等参数，设计并执行"一次只改一个变量"的对照实验；能完成"生成 → 筛选 → 剪辑 → 配乐 → 导出"的完整流水线，并对失败片段写出可执行的改进方案。
3. **素质：**建立"AI 视频是概率生产、必须大浪淘沙"的理性预期与成本意识；掌握《人工智能生成合成内容标识办法》的要求，养成"发布前标识"的操作习惯，抵制虚假信息，坚持诚信自律。

对应第 4 周 · 3 学时　讲授 1 学时 · 实践 2 学时　主线推进：4–6 段镜头视频

凌晨一点，小舟把第 2 章做好的分镜图"水月"拖进视频工具，输入提示词："月夜，江面如镜，一轮明月从潮水中缓缓升起，镜头缓慢推近。"第一次生成的 5 秒里，第 1 秒很美，第 2 秒月亮抖了一下，第 3 秒江面像被揉皱又摊平——**单看每一帧都还算过得去，连起来放却"跳"得像信号不好的老电视**。她一共生成了 7 次，只有 2 次能看，最后剪进片子的那 1 段，是第 6 次生成的。这一章要讲清楚三件事：视频为什么比图像难一个量级（原理）；运动幅度、时长、分辨率、种子这些旋钮各自在管什么（参数）；以及怎么把"抽卡"变成一条可控的流水线，从几张分镜图做出 4–6 段能进成片的镜头（实操）。学完这一章，你手里就集齐了本书前半程最重要的两块拼图。

图 3-1　第 3 章的内容地图。从"数据立方体"到"成片"，本章走完一条完整的镜头生产线：先讲原理（为什么难），再讲控制（怎么让它听话），最后讲工程（怎么把它做出来、做得起、做得合规）。

3.1 视频的时空表征：多出来的那一维

第 2 章结尾，小月交付了 6 张分镜关键帧。但短片是要"动"的：月亮要升起来，江水要流，衣袂要飘。把一张静止的图变成一段动的画面，这件事在技术上叫什么？叫**视频生成**（video generation）。而理解它的第一步，是回答一个看起来很朴素的问题：在计算机眼里，"视频"到底是什么？

◆ 帧：视频的最小单位，是图像

答案的第一半简单得让人失望：**视频就是一叠按顺序排好的图像。**这叠图像里的每一张，叫一帧（frame）；每秒播放多少张，叫**帧率**（frame rate，单位是 fps，frames per second）。电影工业的经典帧率是 24fps，多数生成工具的默认值在 24fps 到 30fps 之间。换算一下就明白"时间"有多贵：**一段 5 秒的 24fps 片段 = 120 帧；10 秒 = 240 帧。**第 2 章你为一帧的好看反复调参数，现在工具要求它在几秒钟里把 120 帧全部画出来，而且画得"前后不像两个人"。

这里就出现了本章第一个关键观念：图像生成与视频生成，不是"多画几张"的关系，而是"多了一整条约束"的关系。图像的目标是"这一帧看起来对"；视频的目标是"每一帧看起来对，而且帧与帧之间的变化看起来也对"。前者是一个二维优化问题，后者是一个带时间轴的优化问题——后面所有的难点（闪烁、漂移、形变、物理错误），都从这条多出来的约束里长出来。

◆ 时空数据立方体：把视频看成"三维"数据

于是，视频在模型眼里有了一个更精确的画像：它不是"一串图"，而是一个**时空数据立方体**——沿着宽（W）、高（H）、时间（T）三个方向同时展开的数据块。一张图像，是这个立方体在 T=1 时的特例；Sora 的技术报告《Video generation models as world simulators》（2024 年 2 月）正是这么处理的：**图像只是"单帧视频"，同一套方法可以处理。**

为什么非要用"立方体"这个词？因为它能解释两件事。第一，**压缩是三维的**：第 2 章的图像模型在潜空间把画面在空间上压小；到了视频，模型的压缩网络（3D 因果 VAE 一类）会连时间方向也一起压缩——把相邻的几帧合并成"一帧的时间表示"。第二，**代价是三维的**：任何一维变大，代价都会同时上涨，而且是叠加性地上涨（下一小节详述）。

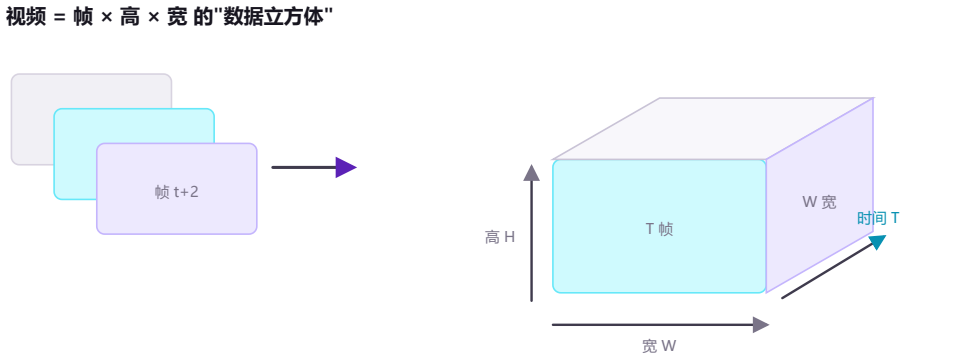


图 3-2　时空数据立方体。视频多出来的那一维是时间。模型要同时学会两件事："每一片切片长什么样"（空间）与"切片之间怎么变"（时间）——第二件事，就是本章的主角。

把视频看成立方体，还有一个立即可用的推论。既然多出来的维度是"时间"，那么"帧率高不高、时长长不长、画面大不大"这三个旋钮转动的代价，就不是相加、而是相乘。举个具体对照：一部 1080p、1 分钟的影片，第三方测算给出的视觉词元数量级是**百万级**（约 100 万到 180 万，随压缩比、patch 大小与采样步数而变，属估算而非官方数字）；而一个更小的公开对照是，模型 Mochi 1 在 480p、约 5.4 秒时，官方模型卡标出的视觉词元数是 **44,520 个。**从"5 秒 480p"到"1 分钟 1080p"，画面并没有变复杂，但模型要处理的"格子"多了几十倍——而注意力机制的计算量是随格子数**平方**上涨的。

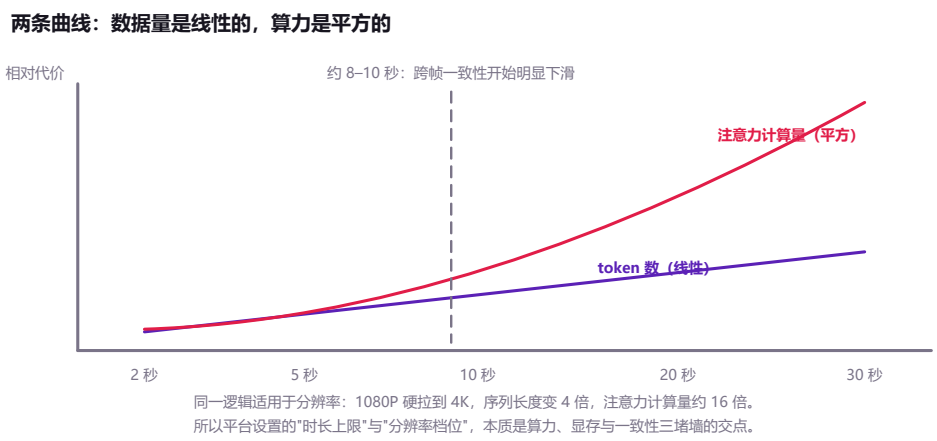


图 3-3　两条增长曲线。这是本章所有"参数语义"的总开关：时长与分辨率每加一档，代价都在平方曲线上往上跳一格。理解这张图，你就明白为什么"更长、更大"的需求总是撞墙。

◆ 难在哪里：三重代价

把上面几段收拢成一句话：视频比图像难，难在"多了一维"带来了三重代价。第一重是**数据代价**：帧数 × 画幅，格子总数成倍增加。第二重是**连续性代价**：模型不仅要生成"好看的帧"，还要生成"帧与帧之间正确的关系"——月亮的位置要在时间上平滑移动，而不是每帧独立地"重新画一个月亮"。第三重是**累积代价**：一帧里的小误差，会在生成下一帧时被当作"事实"接过去，一环带一环地放大（3.3 节整节都在讲这件事）。

你的日常类比是动画。手绘动画师画一个 5 秒的动作要画 120 张原画与中间画，但动画师手里有一条"动作线"：他清楚地知道第 30 帧的手臂应该在哪个位置，因为第 29 帧在那儿。**而 AI 视频模型生成第 30 帧时，它没有"手感"，只有一个学出来的统计直觉**——这既是它惊人高效的原因，也是它时不时"手滑"的原因。第 1 章讲过大语言模型"没有记忆"，视频模型同样如此：它的"记忆"就是你喂给它的条件（首帧、参考图、上一段），而不是它"记得"自己刚才画了什么。

∫ 深入一格 · 为什么"多一维"就等于"平方"　选学 · 数学直觉

模型处理视频时，先把立方体切成小块（时空 patch），每块变成一个词元。设词元数为 N，则 N 大致与"时长 × 宽 × 高"成正比——把时长翻倍，N 就翻倍。

麻烦出在注意力上。第 1 章说过，注意力是"每个词元都看一圈其他词元并加权平均"，因此它的计算量是"词元数 × 词元数"，也就是 N^2。N 翻倍，计算量变 4 倍；分辨率从 1080P 提到 4K（像素 4 倍），计算量约 16 倍。

$$N \propto T \times H \times W \quad \Rightarrow \quad \text{注意力计算量} \propto N^2 \propto (T \times H \times W)^2$$

这是最"朴素"的算法代价。工程上有大量手段把指数压回来：潜空间压缩（先变小再算）、稀疏注意力（只连重要的对）、分块与序列并行（把大矩阵拆开算）——但"平方"这个底色始终在，这就是时长与分辨率有上限的根本原因。

一句话记住这条链：**时间维 → 格子更多 → 平方更贵 → 平台给你设上限。**下次看到某个工具的"单次最长 30 秒"，你就知道那不是产品经理随手写的数字。

表 3-1　图像与视频：多出来的那一维带来了什么

对比项	图像生成（第 2 章）	视频生成（本章）	对学生意味着什么
数据结构	一张画：宽 × 高	一个立方体：宽 × 高 × 时间	视频的产物天然要按"秒"来规划，按"帧"来估算
最小生成单位	1 张图	一段片段（常见 4–10 秒）	分镜要按"镜"组织，一个镜一段，别指望一次生成整片
核心难点	构图、风格、细节、文字	再加"时间一致性"与"物理合理性"	画面好看只是及格线，动得合理才是及格线以上
算力增长	随像素数超线性（"时间墙"）	随像素数 × 帧数平方增长（"平方墙"）	分辨率、时长每一项加码，等待时间与花费都跳档
可控手段	提示词、参考图、局部重绘	再加：首尾帧、参考视频、运镜描述、延长与重绘	条件越多，锚点越强，"抽卡"命中率越高
成功率的现实	尚可，一次常能出可用图	约 3 次生成得 1 个可用镜头（见 3.9 节）	预算时间要按"生成次数"算，不是按"成品时长"算

3.2 三条技术路线：视频模型是怎么"长"出来的

知道了"难在哪"，接下来是一个很实际的问题：模型要怎么设计，才能一边画出好看的每一帧，一边保证帧与帧之间是连贯的？过去三年，这个问题的答案不是三个候选方案里挑一个，而是三条路线**逐层叠加**出来的。理解这个"叠加"关系，你就能看懂今天所有工具背后的共同骨架——也能在别人说"某模型是 XX 架构"时，分辨哪句是对的。

◆ 路线 A · 潜空间视频扩散：给图像模型"装上时间"

最自然的起点是：第 2 章的文生图模型已经很强了，能不能直接用？可以——**拿成熟图像扩散模型当底座，把"时间"当成新部件插进去，再拿视频数据微调。**这就是路线 A，技术上叫潜空间视频扩散（latent video diffusion）。第 2 章说过，扩散模型"在压缩过的潜空间里做去噪"；视频扩散只是在这套流水线里增加了处理时间的层。

具体怎么插？两种典型做法最为常见。**一是时间注意力层**：让不同帧在同一个空间位置上"互相看一眼"，从而协调彼此的样子；**二是三维卷积残差块**：把原来只在平面上滑动的卷积核，沿时间方向也滑动一下。学术界的两篇代表工作把这两条路都验证了：Video LDM（论文题名《Align your Latents》，2023 年 CVPR）明确实现了这两类时间混合层，长视频靠"迭代采样 + 上下文引导"拼接，高帧率靠额外的插帧模型补帧；Stable Video Diffusion（SVD）则把时间卷积与交叉注意力层加在 UNet 的每个空间块之后，并强调**联合微调空间参数与时间参数**比只训时间参数效果更好。它的训练是三段式的：先继承 SD 2.1 的图像预训练，再做大规模视频预训练，最后用约 100 万条精选高质量视频微调。

> 🧪 **两个必须澄清的常见误说**
>
> **其一："SVD 用的是 3D VAE。**"不成立。SVD 的时间层在 UNet 去噪器内部，它的 VAE 沿用 SD 2.1 的图像 VAE，只做空间 8× 压缩，**没有时间维压缩。**真正带时间维压缩的 3D（因果）VAE 出现在后续的 CogVideoX、Open-Sora、HunyuanVideo、Wan 等模型里。写论文、做汇报时这个区别要点清：SVD 时代 = 图像 VAE + UNet 时间层；3D 因果 VAE 时代 = 空间与时间一起压缩。
>
> **其二："Sora 用的是分解式时空注意力。**"不成立。Sora 技术报告里说的 spacetime patches（时空小块）是**输入表示**——视频先被压缩成潜表示，再切成小块当作 Transformer 的词元，这与"注意力怎么连接"是两件事。"patch 化"回答"数据怎么切"，"注意力分解"回答"切完之后谁看谁"。把这两件事混为一谈，是最常见的技术表述错误之一。

◆ 路线 B · 时空注意力：在注意力里显式建模时间

路线 A 解决的是"怎么把图像模型变成视频模型"，路线 B 进一步追问：时间关系到底该在哪里、以什么方式被建模？最著名的答案是 AnimateDiff（ICLR 2024）：向一个**冻结的**文生图模型插入一个即插即用的"运动模块"——一个标准的时序 Transformer，沿着时间轴做时间自注意力，在海量视频上学到"通用的运动先验"。它的巧妙之处在于"冻结"：图像模型负责画面质量，运动模块负责让它动起来，两条能力被解耦开来。论文还展示了 MotionLoRA——大约 50 个参考视频就能适配一种新的运动模式（比如某种特定的摇镜），这为后来的"参考视频控制运镜"埋下了伏笔。

随着模型从 UNet 走向 Transformer，"时空注意力"演进出两种形态。**全时空注意力**让所有时空位置两两相连，表达能力最强，代价是复杂度 $O((T{\cdot}H{\cdot}W)^2)$；**分解式时空注意力**（factorized attention）把它拆成两步——先做"帧内空间注意力"（同一帧里像素块互相看），再做"同位置跨帧时间注意力"（同一个位置在不同帧上互相看），复杂度降为 $O(T(HW)^2 + HW{\cdot}T^2)$，显著更省，代价是一致性略逊。Open-Sora 的 STDiT 就是这一思路的代表："帧内空间自注意力 + 同位置跨帧时间注意力"。

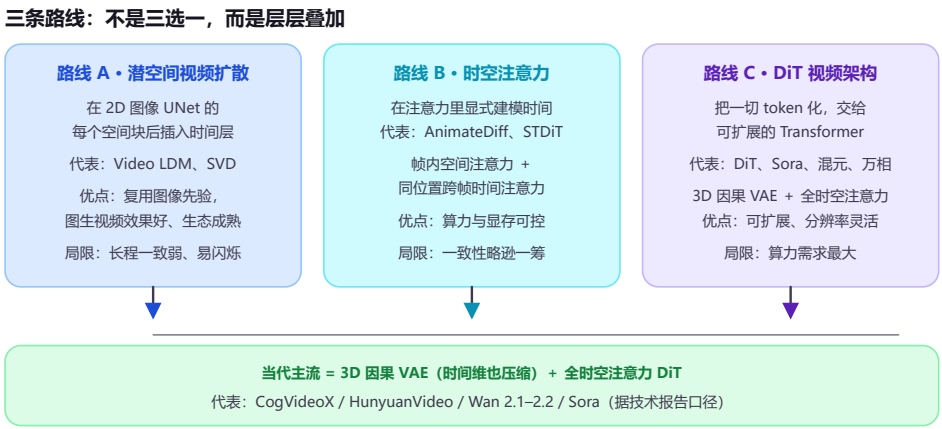


图 3-4　三条技术路线的演进叠加。路线 A 解决"怎么把图像模型变成视频模型"，路线 B 解决"时间在注意力里怎么建模"，路线 C 解决"如何做得更大、更通用"。它们不是互斥流派，而是三层能力叠在同一条链上。

图 3-5　三种注意力连接方式。空间注意力管单帧内的合理性，时间注意力管跨帧的连贯性，全时空注意力把两者合并。图中虚线表示"连接"，连线密度就是算力代价。

◆ 路线 C • DiT 视频架构：把一切交给可扩展的 Transformer

第三条路线的起点是 2022 年的一篇论文：**DiT**（Diffusion Transformer，Peebles 与 Xie）。它做了一个"简单粗暴"的替换——把扩散模型里去噪用的 U-Net 主干换成 Transformer，直接让它跑在潜空间的 patch 上。为什么这件事重要？因为 DiT 论文用一个清晰的指标回答了"换掉之后好在哪"：以**前向计算量（Gflops）**衡量可扩展性，Gflops 越大的模型 FID（生成质量指标，越低越好）持续下降——换句话说，**这条路能"吃算力"**：只要堆规模，质量就跟着涨，而且涨得可预测。论文里最大的模型 DiT-XL/2 在 ImageNet 256×256 上取得了 FID 2.27 的成绩。

后来的一切就顺理成章了：Sora 技术报告明确说明其模型是 diffusion transformer；国内的 CogVideoX、混元视频、万相系列也陆续走到同一条主线上。当代主流的配方可以写成一行公式：3D 因果 VAE（空间与时间一起压缩）+ 全时空注意力 DiT（统一去噪）。各家在这个配方上的取舍，正是你在工具里能感受到的差异来源。

表 3-2　三条技术路线对照（演进叠加，不是三选一）

路线	回答的问题	核心机制	代表工作	优势	代价
A • 潜空间视频扩散	怎么把图像模型变成视频模型	在图像 UNet 的空间块后插入时间注意力层与三维卷积残差块，再在视频上微调	Video LDM（2023）、Stable Video Diffusion	复用强大的图像先验，训练成本低，图生视频效果好，插件生态成熟	图像底座无原生三维时空建模；时长与分辨率为固定小规格；长程一致性弱、易闪烁
B • 时空注意力	时间关系在注意力里怎么建模	时间自注意力（AnimateDiff 式运动模块）；或分解式：帧内空间注意力 + 同位置跨帧时间注意力	AnimateDiff（ICLR 2024）、Open-Sora 的 STDiT	机制清晰、显存与算力可控；运动模块可即插即用，MotionLoRA 约 50 条视频即可适配新运动	分解式的一致性略逊于全时空；全时空则复杂度按平方上涨
C • DiT 视频架构	怎么做得更大、更通用、更可扩展	Transformer 替换 U-Net，在潜空间 patch 上统一去噪；以 3D 因果 VAE 做时空压缩	DiT（2022）、Sora、CogVideoX、HunyuanVideo、Wan	可扩展性强（Gflops 越大质量越好）、原生支持多分辨率与宽高比	算力需求最大，训练与推理成本高

表 3-3　代表性视频模型的技术要点（据论文与官方仓库整理）

模型	年份	核心机制要点
Sora	2024	时空小块（spacetime patch） + DiT；在视频压缩网络的潜空间里训练与生成；以"原生尺寸"训练而非裁剪缩放
CogVideoX	2024（ICLR 2025）	3D 因果 VAE（压缩比 4×8×8） + Expert Transformer；渐进训练与多分辨率帧打包
HunyuanVideo	2024	参数量 13B+；CausalConv3D 的 3D VAE（时间 4× / 空间 8× / 通道 16×）；DiT 全注意力 + "双流→单流"混合设计；用多模态大模型替代传统文本编码器
Wan 2.1	2025	流匹配；Wan-VAE（3D 因果 VAE，步长 4×8×8，潜通道 16，可对 1080P 长视频编解码）；DiT 用全时空注意力（双向非因果）
Wan 2.2	2025	视频扩散引入混合专家（MoE）：高噪声专家管布局、低噪声专家管细节，按信噪比阈值切换；总计 27B、每步激活 14B；某小型版本在单张 24GB 显卡上 9 分钟生成 5 秒 720p@24fps
HunyuanVideo 1.5	2025-11	轻量化 8.3B；SSTA 稀疏注意力；480p/720p 推理、超分至 1080p；默认 121 帧；不含音频；最低 14GB 显存（开启卸载）

表 3-3 里藏着一个对你有直接影响的细节：**HunyuanVideo 1.5 的"最低 14GB 显存"与"默认 121 帧"**。这是开源模型的路线——权重开放、可本地部署，代价是你得自己准备硬件、自己调参数；而网页版工具把这一切藏起来，你只看得到"档位"和"排队时间"。两条路的取舍很清楚：要"可控与免费"，去本地部署；要"省事与稳定"，用网页工具。本书的实操以网页工具为主，但你应该知道本地生态的存在——它是"工具会涨价、会关停"时的退路（3.7 节会讲一个真实的关停案例）。

◆ 算力不够怎么办：四类"降本"手段

三条路线都指向同一个现实：算力是瓶颈。学界与工业界围绕"怎么在有限的算力里生成更好的视频"做了大量工作，可归纳为四类，它们也解释了为什么近两年"同样时长、同样分辨率"的等待时间在缓慢下降：

1. **潜空间压缩。**3D 因果 VAE 在时间维上再压 4 倍（空间 8 倍），等于把立方体的"格子数"直接砍掉一截。这相当于视频编码里的"帧间差分"思路：相邻帧相似的部分不必重复存储。
2. **稀疏注意力。**Sparse VideoGen（ICML 2025）提出一种免训练方案：在线判断每个注意力头是"空间头"还是"时间头"，只保留必要的注意力连接，在 CogVideoX-v1.5 上最高实现 2.28×、在 HunyuanVideo 上最高 2.33× 的端到端加速，且不损伤质量。
3. **蒸馏加速。**CausVid（CVPR 2025）把一个 50 步的双向模型蒸馏成 4 步的自回归生成器，单卡实现 9.4 FPS 的流式生成；后续的 Self-Forcing 改为"训练时就自回归 rollout + 滚动缓存"，把"无限长生成"变成可能。

4 **分块与并行。**把大矩阵拆成瓦片分块推理，把序列切段交给多卡（Ring Attention、Ulysses 等序列并行方案），再配合 token 级裁剪——这是"用工程换算力"的常规操作。

∫ 深入一格 · 时间维"压缩"到底压掉了什么 选学 · 原理

你可能会想：时间维压缩 4 倍，会不会把"动"这件事也压没了？其实不会——关键在于压缩的是**冗余**。相邻两帧 95% 的内容是一样的（背景、光影、色块），真正变化的只是"动"的那一小块。3D 因果 VAE 做的事情，可以类比成视频文件格式里的帧间预测：**只记录"这一帧相对上一帧变了什么"**，而不是把每一帧都完整存一遍。

"因果"（causal）两个字也很关键：它意味着压缩时**只看过去、不看未来**——第 t 帧的表示只能依赖 1..t 帧，不能偷看第 t+1 帧。这个约束让模型可以流式地生成（一边拍一边压），代价是表达能力略受限制。这也是"全局去噪"与"自回归生成"两种范式的分野所在：前者一致性最好但长度固定，后者可以无限长但会漂移（3.3 节细说）。

图 3-6　流体镜头：时间一致性的压力测试（视频）。水、烟、火这类流体的每一帧都在变形，是「帧间不一致」最容易暴露的对象——请重点盯**浪花的形状**：它在连续几帧里是否保持可信的演化，还是像换了张脸一样突然重排？把这条片子和静态分镜放在一起看，你就明白为什么第 3 章要把「时间一致性」单独讲一节。 网页版可播放 AI 生成

3.3 时间一致性：为什么"误差"会自己长大

现在把镜头拉近到本章最核心的难点。小舟那段"月亮抖了一下、江面被揉皱"的视频，专业术语叫**时间一致性**（temporal consistency）出了问题。权威综述《Spatiotemporal Consistency in Video Generation: A Survey》把这件事形式化成"在时空分布上做序列采样"，并把它拆成两个层面：

- **空间一致性**：同一个东西在画面的不同位置、不同方位上看起来还是"同一个东西"——身份、布局、风格、光照都稳定。
- **时间一致性**：运动平滑、状态转移合法、闪烁少、动作语义忠实——前一秒发生的事，与后一秒发生的事在逻辑与物理上接得上。

综述给出的三条根源非常值得抄进你的笔记：**局部误差会累积**、**长程依赖计算昂贵**、**单帧质量好不等于动态连贯**。翻译成创作现场的语言就是：**你可以把每一帧都做得很好看，但错误会像滚雪球一样长大，而要盯着"很久以前的那一帧"保持不偏，代价又高得离谱。**

◆ 两种典型的"时间崩溃"：闪烁与漂移

时间一致性问题在作品里的表现，主要分成两类，它们的症状与成因都不一样，处方自然也完全不同。

第一类：闪烁（flicker）。亮度像有人在打信号灯一样忽明忽暗、纹理像水波一样翻动、边缘持续振动、背景帧间突跳。它的成因有四条，都值得你知道名字：**其一**，VAE 潜时间边界的解码跳变——分块生成时，块与块的交界处缺少真实的"过去上下文"，解码出来就会突跳（有研究甚至用"潜边界处的像素差

异"来度量它）；**其二**，训练与推理时的噪声 / 信噪比分布不一致（FreeInit 论文指出，初始噪声的时间一致性差，根源在此）；**其三**，目标本身的冲突——细节合成鼓励帧间像素大幅变化，一致性又要求约束这些变化，因此一致性机制容易"用力过猛"，把画面搞得过度平滑；**其四**，自注意力缺少时序平滑的归纳偏置——一个可量化的观察是：不连贯的视频，其自注意力图的"时间对角线"呈不规则、碎片化形态；稳定运动的视频则对应平滑的带状对角线。

第二类：漂移（drift）。开头几秒还很正常，越往后画面越"不对劲"：颜色偏了、光线变了、人物的脸像被慢慢溶掉、场景不再是原来那个场景。成因同样有三条：**其一**，因果或流式生成中的误差累积——注意力感受野有限，小预测误差随时间传播，内容逐渐偏离全局上下文（有工作用"让参考图跑在前面当'伪最后一帧'"的办法提供持续的锚点）；**其二**，位置编码越界与频谱能量漂移——帧号变大后，与锚帧的距离超出了模型预训练时见过的范围；**其三**，自回归的暴露偏差（exposure bias）——训练时模型看到的是"真实历史帧"，推理时看到的却是"自己生成的帧"，两者分布不一致，错误由此被逐步放大。

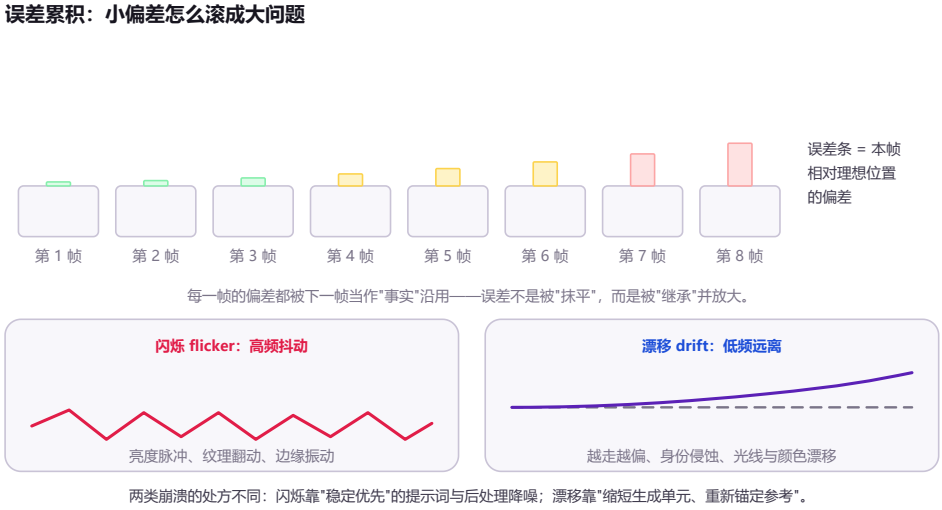


图 3-7　误差累积与两种崩溃。闪烁是"帧间不一致"的高频抖动，漂移是"整体偏离"的低频走样。它们都源于同一个起点：误差被后一帧继承，并且没有足够强的"锚"把它拉回来。

表 3-4　闪烁与漂移：两种时间崩溃的对照

对比项	闪烁（flicker）	漂移（drift）
症状	亮度脉冲、纹理闪烁、边缘振动、背景帧间突跳；观看时"刺眼、像信号不良"	开头正常、越往后越偏；颜色光线变化、人物面部逐渐溶解、场景不再是原来那个场景
时间尺度	帧与帧之间的高频抖动	数秒到数十秒的低频累积
成因（机制层）	VAE 潜时间边界解码跳变；训练-推理的噪声 / 信噪比分布差；"细节 vs 一致性"目标冲突；自注意力缺时序平滑偏置	因果或流式生成的误差累积与感受野受限；位置编码越界与频谱能量漂移；自回归的暴露偏差
提示词侧对策	钉死单一光源；每镜只允许一个镜头运动；负面词加 `no flicker, no warping`；放慢运动、缩短片段	缩短生成单元，靠剪辑而非一次性长生成；每次延长都"从最后帧自然延续"；用参考图重新锚定身份
后处理侧对策	降闪烁（deflicker）、稳定化处理；必要时重生成该片段	重新生成后期片段并与前段做匹配；避免"末帧链式生成"，改用参考生视频

∫ 深入一格 · 为什么"注意力对角线"能诊断视频连贯性　选学 · 原理观察

把模型每一层的注意力图画出来，可以看成一张"词元 × 词元"的热力图。如果按帧顺序排列词元，那么"第 t 帧的某个位置"与"第 t+k 帧的同一位置"之间应该有稳定的关联——把这些关联画在图上，就是一条沿着时间方向的"对角线"。

观察发现：**运动稳定的视频，这条对角线是平滑、连续的带状；不连贯的视频，它呈碎片化、断续的斑点。**这给了一个很有价值的启示——连贯性问题不只是"像素层面看起来不对"，它在模型的内部表示里就有清晰的痕迹。理解了这一点，你就明白为什么研究者常说"一致性要在大模型内部建模"，而不是靠后期"修一修"：**出错的地方在表示层，补救的动作也必须在表示层。**现在已有工作尝试用这类信号做质量检测与自动化筛选，未来很可能变成生成工具的"自动质检"功能。

◆ 更深一层：不止要"连贯"，还要"正确"

时间一致性解决的只是"接得上"，还有一层更难的要求：**接得上是"对"的。**月亮升起该慢慢来，落花该往下落，重物落水该有溅起的水花——这些常识叫物理规律。近两年，"视频生成模型是不是世界模拟器"成了整个行业最热的话题，因为它直接关系到"AI 视频能不能用来做严肃的内容"。

这场讨论有两组证据，都值得你完整地了解——只讲一半都会得出偏颇的结论。

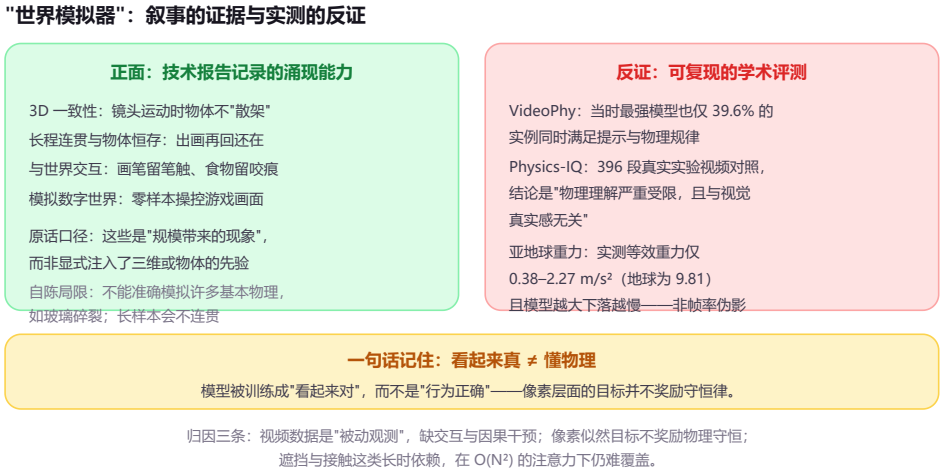


图 3-8　世界模拟器叙事与它的反证。一边是技术报告记录的涌现能力，一边是可复现的学术评测。把两面都摆出来，才是一种诚实的理解方式：模型确实"学到了"很多物理直觉，但离"可靠地模拟世界"还有距离。

对创作者的实践含义有三条。**第一，别把精确物理交给模型**：碰撞、碎裂、软体、液体的精细演出，要么改用实拍，要么走传统特效，要么用短片段 + 剪辑绕开——让模型去承担它做不到的事，就是给自己制造返工。**第二，误差随帧数累积，所以"短片段拼接"胜过"一次性长生成"**：同样 20 秒的内容，四段 5 秒拼起来，往往比一次生成 20 秒更稳。**第三，物理准确性不会随模型变大自动变好**：评测显示更大的模型在"下落速度"上反而更慢——所以"换更强的模型"不总是解决方案，先看问题属于哪一类。补一句学术侧的好消息：有研究发现，只用约 100 段单球下落片段训练一个轻量适配模块，就能把某个 5B 模型的"等效重力"从 1.81 提升到 6.43 m/s²（约地球的 65%），并且能零样本泛化到双球与斜面场景——"物理直觉"是可以被低成本补强的，这为将来的工具留下了想象空间。

3.4 让画面自己"出声"：音视频联合生成

2025 年 5 月 20 日，Google 在 I/O 大会上发布 Veo 3，官方博客的原话是它"首次能够生成带音频的视频"，举的例子包括街道的交通噪声、公园的鸟鸣，**甚至角色之间的对话**；同段话里还有一句值得注意的表述——"在文本与图像提示、真实世界物理以及准确的口型同步上，Veo 3 都表现出色"。行业普遍把这一天视作"原生音频"的转折点。四个月后，OpenAI 发布 Sora 2（2025 年 9 月 30 日），宣布它可以生成**与画面同步的对白与音效**（初代 Sora 不能），并改进了物理真实感（动量、浮力、物体恒存），同时诚实地承认"改进了，但不完美"。

国内跟进得极快：可灵 4.0（2026 年 9 月底至 10 月）给出高品质双通道立体声与口型同步，覆盖中、英、日、韩、西班牙语等多语种及多种方言口音；MiniMax H3 提供双声道音频，多轨输出覆盖音乐、音效、配音，官方博客也坦承仍存在"偶发音频失真"；字节的豆包 Seedance 2.0（2026 年 2 月 12 日发布）基于"统一多模态音视频联合生成架构"，提供双声道音频，支持 15 秒高质量多镜头音视频输出。开源阵营同样热闹：Ovi 用"双塔 DiT + 分块跨模态融合"开源了 11B 权重，5 秒 24fps、720×720；百度的 NAVA 基于 Wan2.2 骨干做"先对齐再融合"，这个 6.3B 的联合模型输出双声道音频（场景声与语音联合去噪，不需要事后用声码器对齐），在 8 卡并行下可生成 720p、约 1 分钟的素材；Lightricks 的 LTX-2.3 是 22B 的联合音视频基础模型，最高 4K/50fps/20 秒，权重开源。

为什么"让画面出声"这么难？因为它要求模型同时掌握两套完全不同的信号：画面是空间上高度结构化的二维网格，声音是在时间上一维流动的波形。把两者放进同一个生成过程，本质上是在解决一个"同步"问题——而这个同步有两种完全不同的做法。

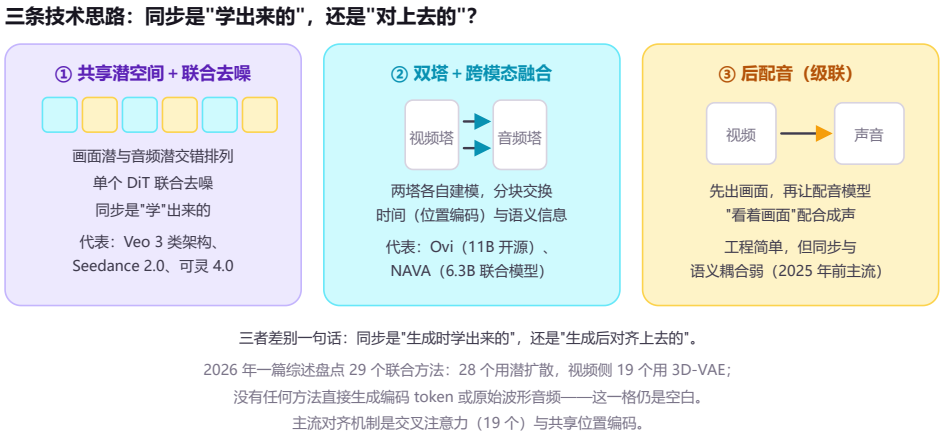


图 3-9　音视频联合生成的三条思路。从"先生成后配音"到"画面与声音一起长出来"，同步质量一路上升，代价是模型与数据的复杂度也在上升。创作实践中，两者会长期共存。

> **📌 选工具时，"有没有声音"是一等一的筛选条件**
>
> 不同工具对音频的支持差异极大，而且都写在参数里、不写在宣传语里：通义万相 3.0 的原生音频参数默认开启，且官方明确"开关声音的价格相同"；Vidu 从 Q3 起支持背景音乐与音效参数，但默认关闭；腾讯混元视频 1.5 明确不含音频。Hailuo 的接口里还有"驱动音频"（driving audio）一类输入，可以让生成的口型跟着你给的一段音频走。
>
> 操作建议：**先定"这一镜要不要出声"**，再选模型。要现场对白与音效的，选原生联合生成的模型；只要配乐与旁白的，第 1 章的 TTS 与音乐工具链更便宜也更可控。另外提醒一句：联合生成的音频仍然需要人工试听把关——厂商自己也承认存在"偶发音频失真"这类问题。

表 3-5　音视频联合生成的三条思路对照

思路	做法	同步质量	工程复杂度	代表
共享潜空间 + 联合去噪	把画面潜表示与音频潜表示交错排进同一个序列，用单个 DiT 联合去噪，再联合解码	最高。同步是"学出来"的——模型在训练中学会了"这种画面配这种声音"	最高，需要大规模音视频配对数据与联合训练	Veo 3 类架构；Seedance 2.0 的统一多模态架构；可灵 4.0 双通道立体声
双塔 + 跨模态注意力	视频与音频各用一个塔（或各自的主干）建模，再通过跨模态注意力与共享位置编码交换信息	较高。两塔能力可独立迭代，融合层负责对齐	中等偏高，但比联合训练灵活	Ovi（twin-DiT 分块融合，11B 开源）；NAVA（"先对齐再融合"，6.3B，基于 Wan2.2 骨干）
后配音（级联）	先生成无声视频，再由音频生成模型"看着画面"配音配乐配效	较弱。同步靠事后对齐，语义耦合有限	最低，模块化、易替换	2025 年前的主流做法；今天的通用流程中仍广泛用于旁白与配乐

3.5 条件控制：你想让它怎么动，就怎么"喂"它

原理讲完，进入你每天都在用的层面：**怎么让模型生成你想要的画面与运动。**这一节的内容，可以把第 1 章的"提示词四要素"与第 2 章的"参考图"两套本事合并升级——视频比图像多出一整套条件系统，因为除了"画面要什么"，你还得交代"它怎么动"。

按"你给它什么"来分，视频生成的条件有五类：**只给文字**（文生视频）、**给一张起始图**（图生视频）、**给首帧与尾帧**（首尾帧插值）、**给参考图或参考视频**（主体与风格一致性）、**给运动指令**（轨迹、相机运动或运镜描述）。它们可以叠加，但注意：部分平台明确规定了某些条件是互斥的——例如万相 3.0 的文档写明"首帧/尾帧"与"参考图/参考视频/参考音频"不能混用；海螺同样规定首尾帧与参考类型互斥。选条件之前先看文档，能省掉一整轮无效生成。

图 3-10　四种条件方式的对照。从"只给文字"到"给参考素材"，可控性逐级上升。主线案例的做法很典型：**先图生视频（用分镜图当首帧），再叠加参考图与运镜描述**——把"长什么样"交给图，把"怎么动"交给文字。

文生视频与图生视频：先有锚，再谈动

文生视频（text-to-video）只给文字，是自由度最高的方式，也最适合环境、氛围、空镜这类"没有明确主角"的镜头。它的代价在第 2 章你已经体会过了：**同一段提示词跑三次，可能得到三个不同的世界。**视频版本更麻烦——不仅"世界"会变，连"变的方式"也会变。所以文生视频的正确定位是：**用来探索与找感觉，而不是用来做需要接戏的镜头。**

图生视频（image-to-video）是本书主线的主力做法：把第 2 章做好的分镜关键帧当作**首帧**传进去，再写"接下来发生什么"。好处立竿见影——身份、构图、色调都被这张图"钉"住了，模型要做的只是"让它动起来"。写提示词时有一条纪律务必记住：**不要再重复描述图里已经明确的信息。**你上传了模特图，就不必再写"白袍男子、长发束冠"——写了反而会与图"抢方向盘"，这是第 2 章"两个事实来源打架"那条纪律在视频上的同一条延伸。图里已经有的，交给图；要变化的，才写进提示词。

两者配合的典型节奏是：**文生视频找氛围，图生视频定成片。**小舟的做法就是这样：先用文生视频生成 3–5 版"月升"的感觉，圈出最喜欢的那一种"运动气质"，回到第 2 章的关键帧上用图生视频把镜头做实。

图 3-11　图生视频的「第一帧」从哪来。这一段镜头不是从零生成的：它把第 2 章产出的分镜图（上图）作为首帧喂给视频模型，让画面「从这个状态开始动」。这就是「流水线」的实际含义——上一章的资产，是这一章的锚点。 AI 生成

首尾帧插值：告诉模型"从哪来、到哪去"

图生视频只给起点，**首尾帧**（first/last frame）插值则把终点也定下来：你上传第一帧与最后一帧，让模型补齐中间的过程。这项能力在 2025 年前后集中落地——Luma 的 Ray2 在 2025 年 3 月上线时官方支持 Start/End Keyframes（同时上线延长与循环功能，5 秒可延伸至 30 秒），其 API 的关键帧对象既接受图像，也接受"某次生成的编号"；可灵提供 `first_frame` 与 `end_frame` 参数，但**不支持只给尾帧**，且首尾帧模式会自动继承输入图的比例；通义万相 3.0 则同时提供首帧与尾帧参数，并在文档里注明它与参考素材类条件互斥。

时长方面要留意一个有教学价值的差别：**首尾帧模式往往有独立的时长限制。**可灵某版本接口的时长枚举为 3–10 秒，而文生与首帧图生只支持 5 秒与 10 秒；万相 2.2 的首尾帧接口更严格——时长固定为 5 秒，不可修改。这就是为什么"单次时长"和"累计上限"必须分开看：同一个平台里，不同模式的上限都不一样。

首尾帧的官方提示词写法有一个很实用的模板（可灵用户指南的例子）："Take [@Image1] as the start frame, take [@Image2] as the end frame, [然后描述变化]"。这个模板揭示的正是首尾帧的核心逻辑：**首尾两帧定义"边界条件"，提示词只负责描述"中间该怎么走"。**此外，官方的已知局限也要记牢：起止两帧差异过大时，模型会自行编造过渡，且不一定与你的意图一致；因此首尾帧最擅长的是"几秒钟的主镜头运动"，而不是"跨越两个场景的大跳变"。

图 3-12　首尾帧的有效配对与无效配对。首尾帧不是"随便两张图"，而是一组必须成立的"边界条件"。判据的实质是：**两帧要在同一个空间、同一套光照里，讲述同一个动作的两个状态。**

参考图与主体一致性：把"是谁"交给图

如果你要把同一个角色放进多个镜头，参考图就是最重要的条件。第 2 章已经建立了"参考图解决'长得像'、提示词负责'在做什么'"的分工，视频侧只是把它变成一套更复杂的**参考素材系统**：

- **可灵 Elements（元素）**：每个"元素"至少需要 2 张参考图（1 张主图 + 补充图），最多 4 张；接口上对应主图与补充图两个字段。使用时在提示词里用 `@Element1` 这样的语义标签引用，形成"参考图—语义标签"的绑定——这一点很有用：它让你能像写剧本一样写提示词（"@Element1 抬头望向月亮"），而不用反复描述长相。Elements 不限于人物，动物、道具、服装、场景都可以。数量上限方面，某版本接口规定：有参考视频时参考图不超过 4 张，无参考视频时不超过 7 张；参考主体最多 3 个；首尾帧模式下仅部分模型支持 Elements。
- **Vidu 的参考生视频**：多数模型的参考图数量为 1～7 张；其中 `viduq2-pro` 还能混传视频（此时图片降为 1～4 张、视频 1～2 个）。它还提供面向行业的专用模型（广告、剧情、混剪），本质是把"参考素材的使用方式"预设成了模板。
- **万相 3.0 的参考体系**：参考图最多 10 张、参考视频最多 5 段（总时长不超过 15 秒）、参考音频最多 5 段（总时长不超过 15 秒），可自由组合；提示词里用"图1""视频1""音频1"来标注对应的素材。

> **🧪 机制注 · "它怎么记住这张脸"**
>
> 你可能会好奇：模型是怎么做到"看了两张照片就能在视频里保持这张脸"的？公开资料能给到的诚实答案是有限的：可灵团队在访谈中提到，人物特征的保持来自**模型内置的指令理解**——先把照片中的关键特征解析出来，再把这些特征注入视频生成任务。至于参考编码网络的具体结构与细节，公开资料并未披露。
>
> 本书的态度是：**把结论留在能证实的层面。**你知道"参考图会被编码成某种条件、注入生成过程"就够了（第 2 章讲的三种条件注入方式——通道拼接、交叉注意力、自适应归一化——在这里通用）；至于"第几层注入、用了多少维向量"这类细节，不要凭空推测，更不要写进你的技术说明里。这是学术诚信的一部分，也是 3.13 节思政要展开的话题。

运动控制：让镜头"按你说的走"

比"是谁"更难控制的是"怎么动"。学术上主要有三条思路：

1. **相机与物体解耦控制。**MotionCtrl（SIGGRAPH 2024）在既有视频模型上加了两个适配模块：一个负责把相机位姿序列"时间性地"注入（CMCM），另一个负责把物体的逐帧位置"空间性地"注入卷积层（OMCM），从而实现相机运动与物体运动的**解耦**——这在创作上意义很大：你想要"镜头绕圈、人不动"和"镜头不动、人走动"是两种完全不同的效果。

2. **位姿参数编码。**CameraCtrl 是即插即用的相机位姿控制模块，它的关键设计是用 Plücker 嵌入（每个像素 6 维）来表达光线方向，而不是直接把相机参数喂给模型——原因是旋转矩阵有正交约束、平移尺度又无界，直接学容易"失配"。相机特征还会与图像潜特征逐像素相加，注入到每个时间注意力模块的第一层。

3. **轨迹控制。**DragNUWA 允许文本、图像、轨迹三种输入同时使用，从语义、空间、时间三个视角做细粒度控制；Tora（CVPR 2025，阿里）是首个轨迹导向的 DiT 框架，它的做法是把轨迹转成帧间位移图、编码成分层的时空运动 patch，再用自适应归一化注入并"零初始化"，以最小代价保留底座能力，支持到 204 帧、720p。

> ⚠️ **工程口径：网页版工具的"运镜控制"其实不给你位姿参数**
>
> 上面那些是研究层面的机制。落到你手上的网页工具，多数**并不暴露相机位姿参数**——你能用的只有两条路：**用提示词术语描述运镜**，或者**上传一段参考视频让模型复刻它的运镜**（例如即梦的参考视频控制就明确可以传递运镜速度、人物轨迹、动作编排、多人交互）。理解了这一层，你就不会在界面上白找"相机角度"的滑块了——在当下的工具里，运镜主要是"写出来"或"示范出来"的，而不是"设出来"的。

◆ 把镜头语言翻译成提示词：第 1 章的"分镜表"到这里才真正落地

第 1 章做分镜时，你已经在每个镜头后面标注过景别与机位运动；第 2 章把它们写进了画面任务表。现在到了关键时刻：这些标注要变成模型听得懂的句子。这一小节是全章最具"手艺"的部分——它决定了你的镜头是"有导演感"，还是"像一张会动的图片"。

先复习三个基础概念，它们决定了你的镜头能不能"接得上"。**景别**指被摄主体在画面里呈现的范围大小，以正常成年人为参照分为远景、全景、中景、近景、特写五个基本档（另有更细的大远景、大特写）。决定景别的是"机位与主体的距离"和"镜头焦距"两个因素——**它是相对概念，不是绝对刻度**。其中**中景被称为"万能镜头"**，因为它既能看到环境又能看到人的活动，与任何景别相接都不突兀；而**远景与特写这两个"两极镜头"一般不能直接相接**，不符合人眼看东西的规律。**轴线**则是被摄对象的视线方向、运动方向与彼此关系所形成的一条假定的直线；拍摄时机位要在轴线同一侧（180 度内），否则就会出现"越轴"，造成方向与空间的混乱。**越轴不是绝对禁忌**——不少导演有意越轴以制造特殊效果（黑泽明《罗生门》的奔跑段落就是经典例子），但那是"故意为之"，不是"不小心为之"。

更重要的是四条"组接硬规则"，它们同样适用于 AI 生成的多镜头序列：

1. **景别级差原则。**相邻两个镜头的景别要有明显区别，一般保持两档以上。可以把层级写成：远景(1) ‹ 全景(2) ‹ 中景(3) ‹ 近景(4) ‹ 特写(5)。相同景别、角度变化又小于 30 度，必须回去改分镜；跨两级以上则允许、无角度限制。
2. **30 度原则。**切换同一场景的不同机位时，前后两个镜头的拍摄角度差应大于 30 度。角度差太小，观众感知到的不是"换了个机位"，而是画面在"跳"（跳切感）。
3. **180 度轴线规则。**不要越轴。这条在 AI 生成里比实拍更容易被违反——因为**每一镜都是独立生成的，模型并不知道你上一镜把摄影机放在了哪一侧。**
4. **运动衔接规则。**静接静（内容连续、逻辑成立）；动接静（运动趋于趋缓时才可组接）；动接动（运动方向不能相反——推镜头不能接拉镜头，否则会"跳"）。此外还要留意视觉重心（主体在画面中的位置）的衔接。

于是出现了一件很实际的事：**这些规则必须在"写提示词之前"就写进分镜表**，而不是等生成完再回头修补。小舟在主线上给每一镜加了两个新字段——"上一镜景别 / 本镜景别"与"机位在哪一侧"，就是为了避免"两条相邻镜头都是远景、角度还一样"这种一到剪辑台才发现的事故。

接下来是运镜本身。九种基本运镜的写法与用途，请看下面这张九宫格——建议你把它打印出来贴在工位上，写提示词时对照着抄。

图 3-13　九种基本运镜速查。左侧是名称与写法，右侧是运动方向示意。写提示词时的口诀是"**一个镜头、一个主要运动、一个速度词**"——把这张图当作你的运镜字典。

表 3-6　六类条件控制方式速查

方式	你给什么	模型补什么	平台与参数	适用与坑
文生视频	一段描述画面与运动的文字	全部内容：主体、场景、运动、光线	几乎所有工具都支持；通常写在提示词框，配比例、时长、分辨率档位	适合氛围、空镜与探索；主体不稳定，不适合需要接戏的角色镜头
图生视频（首帧）	一张起始图 + 描述"接下来发生什么"	时间的展开：运动、后续画面	可灵、即梦、海螺、Vidu、万相等均支持；首帧就是输入图	主线主力做法；**不要在提示词里重复描述图里已有的信息**
首尾帧插值	首帧 + 尾帧 + 中间过程描述	中间的过渡过程	可灵（不支持仅尾帧，自动继承输入图比例）；万相（与参考素材互斥）；Luma 的 Start/End Keyframes	适合几秒钟的主镜头运动与状态变化；**两帧差异过大时模型自行编造过渡**；注意部分接口时长固定
参考图 / 主体一致性	1–10 张参考图（视平台）	把主体长相与风格锚定下来	可灵 Elements（每元素 2–4 张，提示词用 @Element1 引用）；Vidu 参考生视频（多数 1–7 张）；万相（参考图 ≤10、参考视频 ≤5、参考音频 ≤5）	多镜头项目的必需品；参考图要"同场次同光线"，避免一张微笑一张中性
参考视频	一段参考视频（+ 参考图）	复刻其运镜速度、人物轨迹、动作编排、多人交互	即梦的参考视频控制；Vidu 某版本可混传 1–2 段视频（此时图片降为 1–4 张）	解决"说不清的运镜"；注意参考视频与参考图数量会互相挤占配额
运动控制	轨迹、相机运动，或用文字描述运镜	按指定路径与方向运动	网页工具以提示词描述或参考视频复刻为主；研究侧有轨迹与位姿控制（Tora、CameraCtrl 等）	轨迹近邻或交叠时容易"串轨"；运镜描述遵循"每镜一个主要运动"

把上面这些落成一句可执行的话，就是视频版的提示词通用公式（多家官方文档的写法高度一致）：

主体描述 + 运动方式 + 运动幅度 + 运镜类型 +（景别 · 角度 · 光影 · 质感 · 稳定约束）

百度"蒸汽机"的官方文档给出的骨架就是"主体描述 + 运动方式 + 运动幅度 + 运镜类型"，并要求写明景别（中景/特写/全景/远景/近景）、角度（平视/仰拍/俯拍）与幅度（轻微/中等运动）。可灵的模板更强调结构：主体 + 运动 + 场景 +（电影语言 + 光影 + 氛围）。万相 3.0 的官方提示词指南则把结构拆成：总体描述 + 参考素材引用 + 分镜（起止秒 + 主体 + 场景 + 运动 + 美学控制）+ 台词 + 音效/BGM + 风格情绪 + 负向清单。

还有一种更"导演化"的写法值得单独学，来自一份厂方实战手册：**起幅 → 运动 → 主体关系 → 落幅 → 目的。**它把"这个镜头为什么存在"也写进了提示词里（Starting Frame + Subject Action + Camera Path + Speed + Subject Relationship + Lens + Occlusion + Ending Frame + Story Purpose）。它的价值在于强迫你想清楚"落幅"——如果不写落幅，模型不知道镜头该停在哪里：只写"环绕镜头"，模型可能绕到一半就结束；写成"顺时针 120 度环绕角色，固定半径约 1.5 米，结束于四分之三特写"，你才拿到一个能接下一镜的镜头。

表 3-7　运镜写作的六条纪律（附反例与正例）

纪律	反例	正例 / 替代做法
一个镜头只用一个主要运动 + 一个速度词	"快速推近的同时环绕并升镜"——叠加 pan + orbit + crane，几乎必然产生不稳定运动	"镜头缓慢推近，从中景推进至面部特写"
术语不够，要描述画面变化	只写"希区柯克变焦"——模型认识这个词，但不一定能复现你想要的效果	把稀有术语翻译成"屏幕上可观察到的结果"，如"从前景树叶平滑变焦到背景人物"
专业电影术语一两个即可	一句话堆进五个术语，模型"抓不住重点"	术语 + 一句白话解释；需要堆叠时拆成两个镜头
必须写落幅	只写"环绕镜头"——模型不知道镜头停在哪里	"顺时针 120 度环绕角色，固定半径约 1.5 米，结束于四分之三特写"
必须写镜头与主体的关系	只描述主体，不交代镜头相对主体的位置与距离	"与跑步者同速侧面跟拍，保持全身侧面轮廓"
只描述内容、不写运镜指令	"月夜，江面，一轮明月"——没有运动信息	补上"镜头缓慢推近、明月自潮水升起"——否则结果会"像动态 PPT"

最后补一句"稳定声明"的技巧：万相 3.0 的官方提示词指南把"负向清单"列为提示词结构的一环，你可以照抄这个模板——`no text overlay, no watermark, no flicker, no warping, no extra fingers, no duplicate characters`；也可以在提示词末尾写一句稳定性声明："无闪烁、无重影、无抖动，运镜平稳无晃动"。它不能包治百病，但成本为零，值得养成习惯。

✎ 提示词 · 镜头 V1 · 月升（图生视频）

以 K1 分镜图为首帧。镜头缓慢推近，从远景推进至中景；一轮明月从潮水中缓缓升起，**[江面倒影]**随之被拉长；薄雾在月光下缓慢流动，水面只做轻微波纹起伏。固定机位、单一光源（月光），情绪由静到开阔；镜头结束于月与江面各占一半的构图。无闪烁、无重影、无抖动，运镜平稳。

时长 5 秒　比例 16:9　分辨率 1080P　运镜 缓慢推近　负向词 no flicker, no warping

✎ 提示词 · 镜头 V2 · 水月诗人（图生视频 + 参考图）

以 K2 分镜图为首帧，参考角色表锁定身份。@诗人 静立江畔，衣袂被江风缓缓扬起；**[左手自然下垂]**，头部极缓慢地向月亮方向转动约 15 度；镜头以极慢速度向右横移约一米，保持人物位于画面右侧三分之一。夜景、单一月光光源、青蓝色调；结束于人物侧影与月同框。无额外肢体、无面部形变。

时长 5 秒　比例 16:9　参考图 角色表 · 3 张　运镜 缓慢横移　运动幅度 低

🩺 **案例 3-A** § 小舟的第一段镜头：文生视频 vs 图生视频

背景：主线短片《春江花月夜》的第一个镜头是"月升·潮水"（对应第 2 章的 K1）。小舟手里已经有 K1 的分镜图，但她想先做个对照实验：既然文生视频更自由，会不会比图生视频更好看？

做法：她用同一段提示词（就是上面 V1 那张提示词卡片的正文）跑两组：A 组是纯文生视频，只给文字；B 组是图生视频，把 K1 当首帧。两组各生成 3 次，参数（时长 5 秒、比例 16:9、分辨率一档到底）保持一致——严格遵守"一次只改一个变量"。评估维度是她自己定的三条：① 与分镜表的一致性（构图、色调、意象）；② 运动是否合理（江水流向、月亮升降方向）；③ 有没有闪烁或形变。

结果：A 组的 3 版都很"漂亮"，但三版的构图与色调各不相同，其中一版把月亮放到了画面正中、一版的雾气浓得像火灾现场——与分镜表的"江面占三分之二、月亮贴水升起"对不上；三次里只有一次运动方向是合理的（潮水向内、月亮上升）。B 组的三版构图全部贴合分镜图，色调统一，只有一版在第二秒出现了江面纹理的轻微闪烁。**整体判断：B 组 2 次可用，A 组 1 次勉强可用。**

复盘：①文生视频适合"找感觉"，图生视频适合"做镜头"——这个结论在主线上被反复验证；②图生视频并没有消灭失败，它只是把失败的概率从"构图跑偏"降级成了"运动小瑕疵"，而后者往往可以用"缩短到 4 秒重生成"或后处理解决；③这次对照还留下了一份副产物：小舟把 A 组里"运动气质最好"的那一版的提示词片段（"潮水向内涌、月亮缓慢上升"）抄进了 B 组的提示词，命中率进一步提高——**文生视频是"素材库"，图生视频是"装配线"。**

3.6 参数语义：为什么你的"中等运动"不是我的"中等运动"

第 2 章里你已经体会过图像生成的"参数陷阱"：同一句话里的"高分辨率"，在不同工具里含义可能完全不同。视频参数把这件事放大了一整个维度——因为多出来的时间维，让每个参数都在同时影响**内容、运动、时长、成本**四件事。这一节的目标不是让你背参数值（它们半年就会过期），而是让你掌握一套**跨平台通用的"参数阅读法"**。

先记住三条总则，它们是本节全部内容的压缩包：

1. **量纲不可比。**同一个"运动强度"，在一个平台是 0–10 的整数，在另一个平台是 0–1 的小数，还有平台把它与另一个参数联动。跨平台直接抄参数值，等于"用一个尺子上的读数去另一个尺子上划线"。
2. **参数互相耦合。**时长、分辨率、运动幅度、稳定性、成本这五个变量互相拉扯，"全都要"是不存在的——每次调参本质上是一次取舍。
3. **一次只改一个变量。**做对照实验时，其余参数全部锁死，否则你无法把结果归因到任何单个参数上。这是本课程实践环节的纪律，也是所有实验科学的基本功。

◆ 运动幅度：同一个词，几把不同的尺子

"运动幅度"（也叫运动强度、运动量）控制画面里"动多少"。它是视频独有的参数，也是最能体现"量纲不可比"的典型。下表把几种常见口径并排放——请特别注意量程一列的差异：

表 3-8　"运动幅度"的几种常见口径

平台 / 模型	参数名	量程	参考区间（社区经验，以自己实测为准）
Wan 2.2 图生视频（A14B）	`motion_magnitude`	**0–10**	1–2 几乎不动；3–4 眨眼、微笑、树叶飘动；5–6 转头、手势、云朵流动；7–8 走动、转身、波浪起伏；9–10 剧烈动作
Wan 2.2 文生视频（A5B）	`motion_strength`	**0–1**	静态场景 0.2–0.4；产品展示 0.4–0.6；快速动作（体育、舞蹈）0.7–0.9
Open-Sora	Motion Strength	—	与 Aesthetic、Camera Motion 并列的推理选项，按官方推理脚本调整
Seedance 系列（第三方调优口径）	`motion_strength` + `focus`	两个参数联动	有第三方文章给出过具体组合值，**本教材不引用数值**——只记住"它是联动参数"这件事

说明：参数名与量程来自各模型的参数文档；各档位"对应什么画面变化"属于社区整理的经验口径（凡具体数值都应视为"参考"而非"规格"），不同模型版本之间会漂移。写作业时，应当翻你所用平台的官方文档确认量程，再以自己的实测记录为准——**"会查官方文档"本身就是本章要培养的一项能力。**

比背数值更耐用的是四条**定性规律**，它们跨平台成立，且有厂商知识库或论文级依据：

- **"运动预算（motion budget）"经验法则。**要求多轴同时运动（例如"快速变焦 + 俯仰"同时进行）几乎必然产生特征扭曲；其中**垂直俯仰对细长特征的拉伸尤其严重**——这也是为什么人像怕俯仰、楼宇怕升降的原因之一。每个镜头都有一个看不见的"运动预算"，花超了就要以形变偿付。
- **速度突变诱发形变。**NeurIPS 的一篇论文明确指出：突然的速度变化（abrupt velocity changes）"通常会诱发形变伪影（morphing artifacts）"。所以提示词里写"由慢突然加速"是个危险的愿望，写"匀速渐缓"才是安全区。
- **主体运动幅度不要超过画面宽度的四分之一。**大跨度位移会让中间帧"碎掉"——运动幅度越大，模型要在帧与帧之间补的内容越多，失败概率越高。想让主体大幅移动，正确做法是**用剪辑切分两次小位移**，而不是一次大位移。
- **人像比风景更脆弱。**面部结构的容错度远低于景物纹理：人像镜头建议更低的运动档位、更短的动作链；风景可以适当调高，"风、水、云"这类纹理运动对形变的宽容度更高。

这四条合起来，就是本课程对运动参数的"操作顺序"：**先定"这一镜需要多少运动"，再选略低于直觉的档位，最后检查有没有多轴叠加。**宁可让模型欠一点运动（后期可以补帧加速），也不要让它超支（后期救不回来）。

A · 运动主体：水　　B · 运镜：跟拍

图 3-14　两种「运动」的镜头参考。A 的难点在「水」——流体是时间一致性最容易崩的对象；B 的难点在「相机」——跟拍要求背景的视差变化符合物理。写提示词时，「什么在动」和「镜头怎么动」要分开说清楚。 AI 生成

图 3-15　运动幅度的实测对照（同一首帧、同一随机种子）。两条片子用了完全相同的第一帧与随机种子（20261014），**唯一的变量是提示词**：A 要求「画面完全静止，像一张静止的照片」，B 要求「镜头缓缓向前推进并向右横移、衣摆剧烈飘动」。量化结果：按帧间像素差计算的平均运动量，A 为 **0.78**、B 为 **1.23**（相差约 58%）。这就是「运动幅度」在真实工具里的手感——它不是旋钮上的数字，而是**你用语言描述出来的运动量**；也再次印证本章那句提醒：同一个词，在你和模型之间并不是同一把尺子。 网页版可播放 AI 生成

◆ 时长：视频的时长是"拼接出来的"，不是"生成出来的"

这是本章最值得刻进脑子的一句话。你看到的所有"1 分钟 AI 视频"，都不是一次生成出来的，而是若干 5–10 秒片段（加上若干次延长）剪出来的。为什么单次生成撑不长？四条原因各自独立、叠加作用：

1. **注意力窗口与误差累积。**超过约 **8–10 秒**，模型难以维持跨帧一致性，脸、光、动作开始漂移——这正是我们在 3.3 节讲过的误差累积机制在时长尺度上的表现。
2. **算力与显存的非线性增长。**10 秒 24fps 就是 240 帧；要让第 N 帧与前 N−1 帧都保持一致，代价近似平方增长。有测算称，**15 秒 1080p 的 Seedance 2.0 单次渲染可达约 20 分钟**（二级来源，仅作量级参考）。
3. **长视频训练数据稀缺。**高质量、长时长、少剪辑、主体一致的连续视频素材很难找——模型只在训练分布内可靠，分布的边界就是"长度墙"的位置。
4. **产品与资源约束。**成本、等待时间、内容审核共同推动平台限长；高峰时段甚至可能触发轻量推理分支，导致高频细节丢失。

所以教材给出的实践建议是：**把 4–10 秒当作"基本生成单元"**，把长视频视为"用剪辑与延长拼起来的作品"。这也解释了你马上会在工具表里看到的一个现象——各平台的"单次时长"是 2 秒到 30 秒不等，但"累计上限"能到 2–5 分钟：单次时长是模型的物理限制，累计时长是产品的工程拼接。两者不是一个量纲的东西，永远要分列。

那延长到什么时候该停？下面这组经验数据（二级来源的口径）值得记住，它描绘了一条"衰减链"：

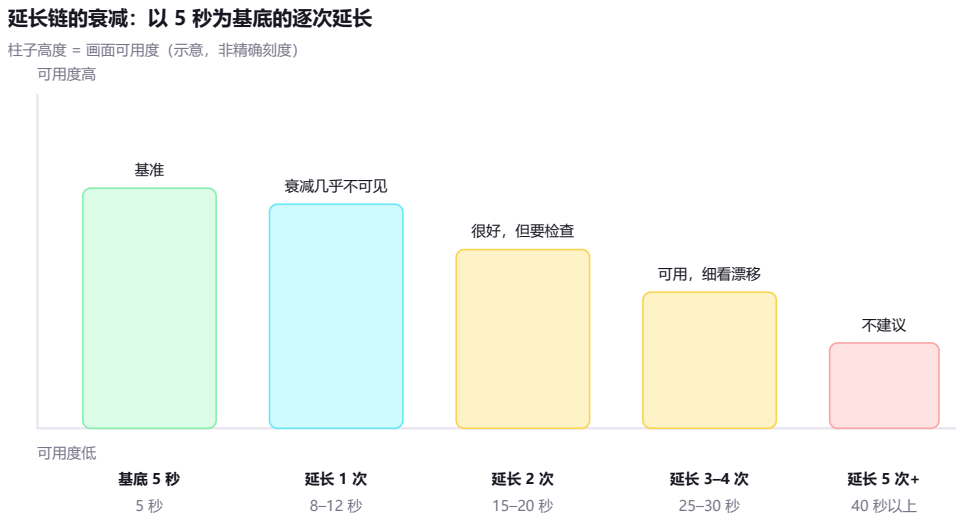


图 3-16　延长链的衰减。1 次延长几乎不可见；2 次延长到 15–20 秒仍很好，但要检查脸和背景；3–4 次到 25–30 秒可用、细看有漂移；**5 次以上、40 秒以上不建议**——场景"身份"会开始被侵蚀。经验来自第三方实测口径，具体平台会上下浮动，但"次数越多、衰减越重"的规律稳定成立。

还有一条更微妙的规律叫**叙事密度上限**：一份基于 MiniMax H3 的边界样本测试认为，**10 秒最多大约容纳 5 个事件节拍**，再往上加就明显"赶戏"。它提醒你：时长参数不只是技术问题，还是"信息设计"问题——不要试图在一个片段里塞进一段完整的故事，那是剪辑的工作。

常见直觉

「参数是各自独立的旋钮：想要 4K 就选 4K，想要 6 秒就选 6 秒，怎么组合是我的自由。」

按这个直觉，你自然会想做一个实验：**固定其他条件，比较"5 秒和 8 秒哪个更清晰"。**

官方规格给出的实情

Veo 3.1（Google 官方文档）：时长有 4 / 6 / 8 秒三档，但 **1080p 与 4K 只支持 8 秒**——想要高分辨率，就必须同时接受最长的那一档。

万相 3.0（阿里云官方文档）：一旦给了输入视频，规则变成**「输入视频总时长 + 输出视频时长 ≤ 30 秒」**——延长不是"再加 30 秒"，而是从同一个总账里扣。

这两个例子指向同一件事：参数不是独立旋钮，而是成组出售的档位。于是"5 秒和 8 秒谁更清晰"这个问题，在 Veo 3.1 上**连实验都做不出来**——你无法在相同分辨率下同时拿到两个时长。

这就是本节的真正结论，也是这一章最该带走的习惯：**参数之间的关系不是"越长越差"或"越大越好"的单向曲线，而是一张到处是耦合的网。**所以下结论之前要做两步：先问**「这两个条件能不能在同一组参数下比较」**（做不出实验的对比，只是猜测）；能比较时，再严格**只改一个变量**——这与第 2 章那两组实测对照图（同种子、单变量）是同一套纪律。**要练的不是记住结论，而是"控制变量再下结论"这个动作。**

◆ 分辨率与等待时间：时间墙，还是显存墙？

视频的分辨率参数和图像一样写着 720P / 1080P / 2K / 4K，但代价结构完全不同：分辨率抬高时，**潜空间序列长度随"宽×高×帧数"增长，而自注意力是 $O(n^2)$**——1080P 拉到 4K，序列长度变 4 倍、注意力计算量约变 **16 倍。**五款主流文生视频模型的实测口径是：把 1080P 配置直接拉到 4K，**推理速度平均下降约 85%，显存占用平均上升约 3.7 倍。**单卡速度律可以记成：端到端耗时 ≈ 固定项 + 步数 × 单步成本，其中**单步成本大致正比于（宽×高×帧数）的 1.5 次方**；因此提速的优先顺序是**"先降分辨率 → 再降步数 → 最后才动注意力内核"。**在 Wan2.2-A14B 的实测中，序列长度约 12 万 token 时，**自注意力占了单步 DiT 延迟的 80% 以上**，显存峰值容易冲过 30GB 而 OOM。

在消费级显卡上，这些规律会直接撞上一堵墙，下面是几组可引用的实测（二级来源，作为"量级感"使用）：

表 3-9　消费级显卡上的"显存墙"实测（本地部署，量级参考）

配置	硬件	耗时	峰值显存
LTX-2.3，97 帧（约 4 秒，带音频）768×512	RTX 5090	约 43.5 秒	24.2 GB
同上，1280×704	RTX 5090	约 49.8 秒	24.2 GB
MiniMax H3，5.9 秒 25 步	RTX 4090 24GB	562 秒	—
同上，8 步 LoRA + int8 量化	RTX 4090 24GB	约 115 秒（提速 4.9×）	—
H3，4K 5 秒	RTX 4090 24GB	约 11 分钟	单个 1472×896 分块

但这里要放一个**重要反例**，它才是这一节最值钱的知识点：如果整条管线的瓶颈在**解码/编码**环节而非扩散模型本身，那么"分辨率越高越慢"就不成立。LTX-2.3 在 RTX 5090 上的一份实测显示：约 43.5 秒的总耗时里，**约 35 秒花在 VAE 解码与 mp4/AAC 编码，DiT 本身只占约 5 秒**——因此把像素数提高 2.5 倍（768×512 → 1280×704），总时间只多出约 14%。同理，图像生成侧也有实测佐证：512/1024/1536/2048 四档分辨率的生成耗时约 0.83/3.18/9.99/23.36 秒，呈超线性（注意力随像素数平方增长），而显存只从约 20.4GB 涨到 28.8GB——被总结为"**时间墙，不是显存墙**"；而**步数的影响则是精确线性的。**

把这两组事实合起来，你就得到了一条比"越高越慢"精确得多的判断法则："更贵"的到底是谁——是扩散计算，还是编解码？先测量，再优化。对网页版工具的用户来说，还有一个现实提醒：平台的等待时间与可选项由平台调度决定，你**看不到显存这一层**，能观察到的只有"档位"和"排队时长"——所以"显存"是给本地部署（开源模型）的学生准备的概念，不是网页版的操作参数。

◆ 种子：可复现的"随机"

扩散模型的每次生成都从一份随机噪声出发，随机数由**种子（seed）**决定。原理一句话：**相同种子 + 相同输入 = 相同输出**；种子留空则每次随机。听起来简单，但它的适用边界必须写清楚：

- **正用法。**① 找到喜欢的效果后**记录种子值**——这是你能"重放"一次成功的最小凭据；② 批量生成时给不同种子，用最少的次数获得最多样的候选（这就是"多抽几张再挑"的成本最优解）；③ 做参数对照实验时**固定种子**，让唯一的变化来自你要测试的那个参数。
- **边界一：平台差异。**种子只在**本地部署或 API 调用**时才稳定可控；多数网页版工具根本不暴露种子——这也是"想做可复现实验的同学应该学一点 API 调用"的原因。
- **边界二：版本漂移。**固定种子通常只在**同模型、同版本、同参数、同平台**下有效；模型一升级，种子就可能失效。你的实验记录里必须写上"平台 + 版本 + 种子"三件套，缺一样就不算可复现。
- **替代方案："身份种子"。**网页版用户拿不到数字种子，但可以拿**一张干净的身份参考图**当角色的 identity seed：一个角色只用一张清晰的设定图，在各个片段中反复复用。这样"复现"的对象就从"噪声"换成了"身份"，同样能显著提高多镜头一致性。

顺便记住图像侧的一个佐证规律，它让你对参数成本的直觉更完整：在图像生成中，**分辨率是超线性的（时间墙），步数是精确线性的**——视频不过是把这两条规律乘以时间维，再叠加一个"帧间一致性"的新约束。

◆ 负面提示词与稳定性声明：零成本的"保险"

最后一组参数没有数字，却值得每次都用：**负面清单**与**稳定性声明。**万相 3.0 的官方提示词指南明确把"负向清单"列为提示词结构的一环，这是它区别于"可选技巧"的证据。教材给的模板是：

> *负面清单：no text overlay, no watermark, no flicker, no warping, no extra fingers, no duplicate characters*
>
> 稳定性声明（写在提示词末尾）：「无闪烁、无重影、无抖动，运镜平稳无晃动」。二者都属于"廉价保险"：不能包治百病，但成本为零、值得养成习惯。注意：负面清单能压制症状，不能修复机制——如果失败原因是身份锚点缺失或运动预算超支，加了负面词也救不回来，此时正确动作是回到 3.5 的条件控制小节去"补条件"。

这一节到这里，参数语义的框架就齐了。请把三条总则再读一遍——它们也是你实践作业里"参数对照实验"的设计依据：**锁变量、记全量（平台＋版本＋种子＋一切非默认参数）、只改一个、多做几组。**下一节我们换一个视角：从"参数"抬到"工具"，看看 2026 年 10 月的工具地图长什么样、又该怎么选。

3.7 工具地图：2026 年 10 月，你手里有什么牌

这一节的全部数据都带一个前缀：**截至 2026 年 10 月。**这不是客套——可灵在一年内走完了 3.0 → 3.0 Omni → 4.0 三个大版本，剪映 2026 年 9 月发布了 Hub，Adobe 2026 年 9 月在 IBC 发布了 Generative Media Tool。**产品规格的半衰期大约是半年。**所以本节的目标是"教你看地图的方法"，而不是"背

下这张地图"：读表时请关注每一列背后的**能力维度**（单次时长 / 分辨率 / 成本口径 / 商用条款），这些维度在换代时相对稳定；具体的数字则以官方文档当场核对为准——"会查官方文档"是本章明确列为学习成果的一项能力。

先看离你最近的一排：国内可以直接访问的工具。

表 3-10　国内 AI 视频工具速查（截至 2026 年 10 月）

工具	版本节点	单次 / 累计时长	分辨率	成本口径	课堂关注的要点
可灵 AI（快手）	4.0（2026-09-28 官宣）；3.0 系列 2026-02-05 全球上线	4.0 单次 30 秒，配合续拍最长约 2 分钟；3.0 为 15 秒	4.0 报 4K 及 1080p 10-bit HDR、21:9；2026-04 起原生 4K 直出	免费每日 66 灵感值（带水印、成片不可商用）；API 720P 约 0.9 元/秒、1080P 约 1.2 元/秒（含音频）	功能最全的一站式：首尾帧、Elements 参考、运动控制、对口型、数字人、多镜头（≤6 分镜）、音画同步、续拍延长
即梦 AI（字节 / Seedance）	Seedance 2.5（2026-07-31 首发接入）	单次 30 秒（原生长镜头）；平台"超长生成模式"最长 3 分钟	4K（即梦官网口径）	网页版积分制，官方仅称提供免费试用积分；API 约 $0.20/秒起（480p 档）	原生音视频联合生成；多模态参考最多 50 项；R2V 参考视频可传递运镜速度、人物轨迹、动作编排；精准局部编辑；**不支持上传含写实真人脸部的素材**
海螺 Hailuo（MiniMax）	MiniMax H3（2026-07-31 发布、08-03 开源）	4–15 秒整数	768P（标准）/ 2K（高清）	官方订阅页：H3 2K 约 ¥0.53–0.78/秒、768P 约 ¥0.31–0.46/秒（按会员档）；年会员低至约 ¥48/月	首尾帧、参考生视频（最多约 9 图 + 3 视频 + 3 音频）、音画一体、动作迁移；官方承认仍存在偶发音频失真等弱项
Vidu（生数科技）	Q4 Preview（2026-10-07 发布）	图生视频 3–16 秒；**视频延长最长可至 5 分钟**	540p / 720p / 1080p / 2K / 4K（2K、4K 为 10-bit）	Q4 首发优惠约 0.09 元/秒起	参考生视频最多约 15 张参考图 + 3 段参考音频；音画同步；**注意：Q4 当前没有文生视频模式**，只有图生与参考生视频
通义万相 视频（阿里）	万相 3.0	最长 30 秒、30fps；有视频输入时"输入 + 输出总时长 ≤30 秒"	默认 1080P，可选 720P / 480P	API：480P 约 0.3、720P 约 0.6、1080P 约 1.2 元/秒	文生、图生、首尾帧、参考生（≤10 图/≤5 视频/≤5 音频）、视频编辑与延长；**原生音频默认开启且开关不加价**；数字人接口
腾讯混元视频	HunyuanVideo 1.5（2025-11 发布权重）	默认 121 帧（约 10 秒 @720p 基准）	480p / 720p 推理，超分至 1080p	**开源、免费**（自备算力）	无音频；最低约 14GB 显存（开启 offload）即可本地部署；许可有地域与规模条款（不适用欧盟/英国/韩国等，详见官方 LICENSE）
PixVerse（爱诗）	V6 / C10 / R2	官方未列单次上限（存疑）	V6 为 1080p + 原生音频	官网称 API 约 $4.80/分钟（基准口径）	角色一致性、对口型、动作模仿、超清；同平台也托管第三方模型

读这张表时请注意三件事。第一，**"单次时长"与"累计上限"分列**：Vidu 的单次是 16 秒，累计却能到 5 分钟——差出来的部分就是"延长 + 拼接"的工程量。第二，**免费额度都附带条件**：带水印、当日清零、不可商用几乎是国产工具免费档的"三件套"，课堂练习可以用，参加比赛和商业发布前必须回到条款页确认。第三，**商用条款是选型的一等公民**：海螺的官方订阅页 FAQ 明确"付费会员期间生成并下载的内容可用于商业用途"；而 Vidu 官方 FAQ 的口径是"是否可商用取决于版本与套餐"。做项目预算时先问"我能不能用"，再问"贵不贵"。

再看国际阵营——多数在中国境内不能直接访问，但你仍需要了解它们，因为它们是"技术风向标"，也是很多国产工具"对标表"的参照系。

表 3-11　国际工具对照（截至 2026 年 10 月；境内多数不能直接访问）

工具	版本状态	单次 / 累计时长	分辨率	价格口径	标志性能力
Sora（OpenAI）**已停运**	停运公告 2026-03-24；据 OpenAI 官方弃用清单，Sora 2 模型与 Videos API 于 **2026-09-24 关闭**（据维基百科，应用于 2026-04-26 停运，此条待官方核实）	曾支持 16 / 20 秒，扩展最多 6 次、总长上限 120 秒	720p；1080p 需 sora-2-pro	曾为 $0.10/秒（标准档）	同步音频、可复用角色、Cameos；它的关停本身成为本章最重要的行业案例（见下文）
Veo（Google）	Veo 3.1（2025-10）+ Fast / Lite 档位	4 / 6 / 8 秒可选；1080p 与 4K 仅 8 秒；延长每次 +7 秒、最多 20 次，最长约 148 秒	720p / 1080p / 4K，24fps	Standard 约 $0.40/秒；Fast 720p 约 $0.10；Lite 720p 约 $0.05（含音频、无免费额度）	原生音频常开、口型同步；首尾帧；参考图 Ingredients（≤3 张）；**所有 Veo 视频带 SynthID 水印**
Runway	Gen-4.5、Gen-4 Turbo、Aleph 2.0、Act-Two，并转售 19 个第三方模型	多数未公布；转售的 Seedance 2.5、WAN 3.0 为 30 秒	视模型而定	统一 credit $0.01/个：Gen-4.5 约 $0.12/秒、Aleph 2.0 约 $0.28/秒	Aleph 2.0 的"上下文视频编辑"（关键帧引导、跨剪辑点编辑）；遮罩擦除；动作捕捉
Luma	Ray3.2（2026-06）；Ray2 与 Dream Machine 已弃用	最长 20 秒 @1080p	1080p；原生 16-bit HDR、可导出 EXR	Plus 约 $30/月、Pro 约 $90/月、Ultra 约 $300/月（含 credits）	多关键帧最多 16 个；Master Reference（角色/产品一致性）；**无原生音频**（音频靠第三方集成）
Kling 国际版（对照）	与国内同源：3.0 / Omni、O1、4.0	3–15 秒；单次多镜头最多 6 个分镜	720P – 4K（30–60fps）	API 标准档约 $0.084/秒（无音频）/ 约 $0.126/秒（含音频）	原生音频（多语言、方言）、动作控制（上传 3–30 秒参考视频迁移全身动作）、数字演员
Midjourney	视频模型 V1（2025-06）	一次 4 段 5 秒；可延长 4 秒 × 最多 4 次 → 最长约 21 秒	各源不一（480p 为多数口径，存疑）	需订阅；**视频成本约为同尺寸图片的 8 倍**	仅图生视频；Auto / Manual 两种运动模式；无音频
Adobe Firefly / Premiere	自有 Firefly 视频模型 + 聚合 Veo 3.1、Kling 3.0、Runway、Luma 等	自有模型 5 秒；多数模型最高 8 秒、部分 15 秒；Premiere Generative Extend 每次最多 +2 秒	自有模型 1080p；生成最高 4K	Firefly 订阅约 $9.99–$199.99/月分档	时间线内直接生成与延长；生成素材带 Content Credentials；**明确不能生成或延长人声对白**；自有模型有 IP 赔偿、合作模型没有

这张表里藏着 2026 年最重要的一条产业信号：**头部阵营正在从"自研单一模型"转向"模型编排（orchestration）"**。Luma 与 Runway 的官网都在转售别家模型（包括可灵、Seedance、WAN），Adobe 把 Google 与 Kling 的模型直接塞进了 Premiere 时间线，Pika 同样如此。对你的直接后果是：越来越难把一个效果归因于"某个模型"，你能控制的其实是"工作流"——同一段提示词在哪几个平台跑、按什么顺序组合、用什么方式回到剪辑台。这也是本课程把 3.8 节"编辑与后期"排在工具表之后的原因。

> **📌 Sora 停运：本章最值钱的一次"失败课"**
>
> 2026 年 3 月 24 日，OpenAI 宣布关停 Sora 业务；据其官方弃用清单，Sora 2 模型与 Videos API 于 2026 年 9 月 24 日正式关闭，官方声明"没有一对一的替代 API（No one-to-one replacement API is available.）"。与之绑定的迪士尼 10 亿美元投资与三年角色授权合作也随之作废。媒体披露的关停经济学（二手口径）：运行成本约 100 万美元/天（WSJ 转引），用户量从约 100 万峰值跌至不足 50 万，应用内总收入估算仅约 210 万美元。
>
> 这件事与"模型能力"无关——Sora 2 在发布时是公认的第一梯队。它说明的是：**模型能力领先 ≠ 产品与生态可持续**。转化为你的一条操作纪律：不要把作业与作品的流程"绑死"在任何一个模型上。素材（分镜图、参考图、提示词、种子记录）永远留在自己的硬盘与笔记里，工具只是随时可替换的一环——这也是教材把"导出与本地保存"写成作业硬性要求的原因。

那么面对这一桌子牌，怎么选？教材给一个"三问决策树"——它不问"哪个模型最强"，只问"你的项目需要什么"：

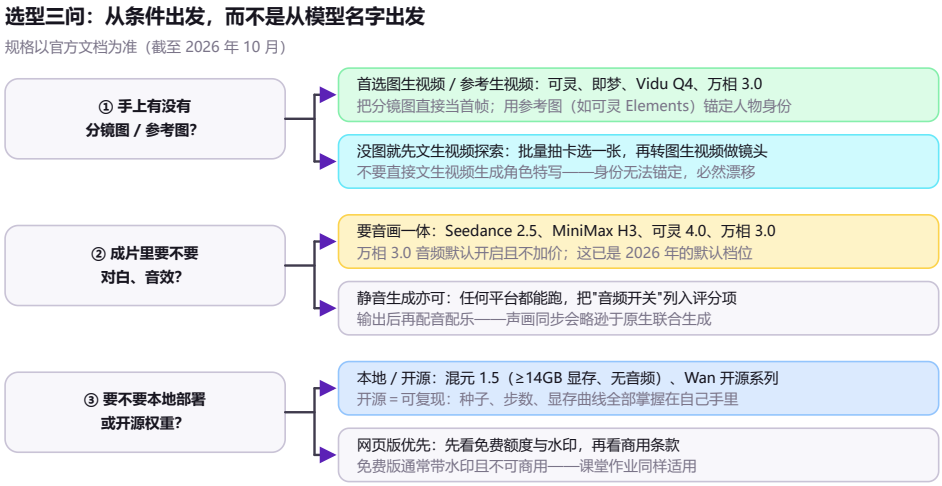


图 3-17　工具选型决策树。三个问题各自独立，按你的项目条件依次回答即可。注意它没有出现"最强模型"这一项——**选型的第一原则是"匹配条件"，第二原则才是"性能排序"**。

如果一定要给"性能排序"，最可操作的客观依据是第三方横评榜——同一套题目、同一套盲测投票，Elo 分数可比。以下是 2026 年 10 月 9 日抓取的快照（仅摘录与课堂相关的前列与国产品牌）：

表 3-12　Artificial Analysis 视频竞技场快照（2026-10-09 抓取，节选）

榜单	前列模型（Elo）
文生视频（含音频，v2.0）	**万相 3.0 第 1（1156）**；Seedance 2.5 第 3（1143）；MiniMax H3 第 4（1137）；可灵 3.0 Omni 1080p Pro 第 15（1015）；Vidu Q3 Pro 第 25（941）
图生视频（v1.0）	**MiniMax H3 Max 第 1（1195）**；MiniMax H3 第 2（1181）；Vidu Q4 Preview 第 3（1179，2026-10 新入榜）；Seedance 2.0 720p 第 5（1176）；万相 3.0 第 7（1164）；可灵 3.0 1080p Pro 第 20（1055）
视频编辑	MiniMax H3 称"全球第一"（媒体转述官方口径）；Seedance 2.5 第 2（1161）

> ⚠️ **榜单的四条使用纪律**
>
> ① **快照日期各不相同**——2026 年 4 月、8 月、9 月的排名不可直接混用；② **AA 的 Elo 与 VBench 的分数不可比**——两套刻度体系不同，同一模型在两处差异可能很大；③ **VBench 从未收录任何 Seedance 模型**，回答不了 2026 年新模型的对比问题；④ 榜单是**人类偏好投票**，反映"好看"而非"物理正确"——请把它和 3.3 节 Physics-IQ 的结论对照着理解：观众觉得好看、模型却可能根本不懂物理，这两件事可以同时成立。

最后补一句关于"生命周期"的提醒。工具会消失：腾讯智影据百度百科口径已于 2025 年 6 月底暂停服务（2026 年是否恢复运营存疑），Sora 也走完了从万众瞩目到关停的全程——而这两个名字在网上至今仍有大量"教程"在流通。**面对快速过期的世界，你的对策有三条**：把"查官方文档"变成本能；把素材与记录存在自己手里；把核心能力练在"方法"（分镜、条件控制、参数实验、剪辑）而不是"按钮位置"上。方法迁移的成本是以天计的，而工具迁移的成本是以小时计的——先练方法，再玩工具。

3.8 编辑与后期：把"抽卡"变成"作品"

生成只是原材料的采购。一部成片的诞生，走的是这条流水线：**生成 → 筛选 → 编辑 → 配乐 → 导出**。前面几节解决的是第一环节；这一节把后四个环节补齐——它们不仅是"技术活"，更是你和"抽卡玩家"之间真正的分水岭：会生成的人很多，会让素材"听话"的人很少。

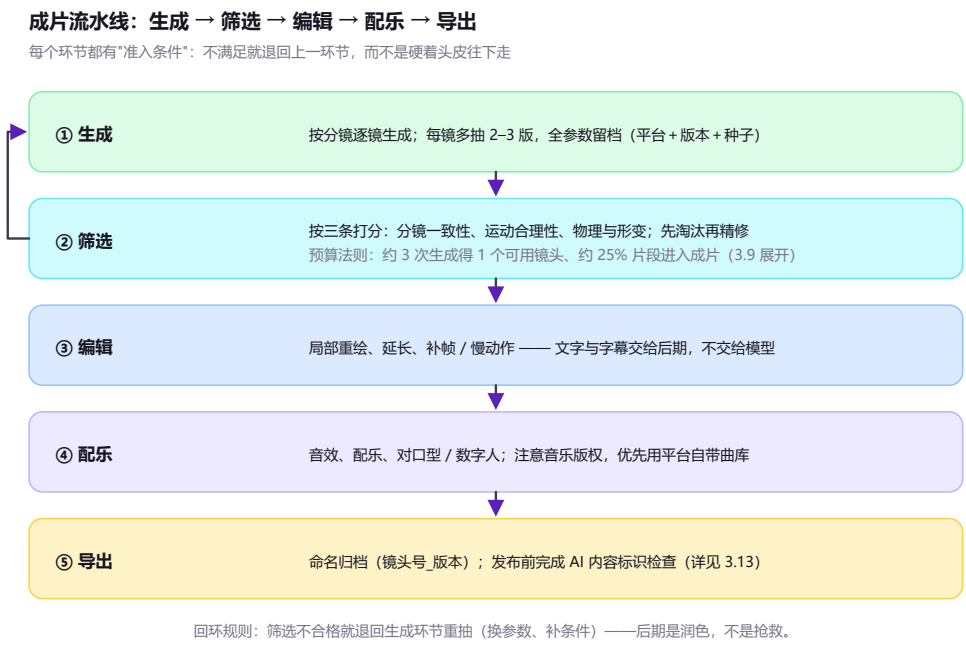


图 3-18　成片流水线。注意左侧那条回环箭头：**筛选不合格时的正确动作是"回到生成"（换参数、补条件、重抽），而不是带着瑕疵进入编辑**——后期是"润色"环节，不是"抢救"环节。流程里的每次退回都比硬凑便宜。

把这五个环节需要的能力集中起来，就得到下面这张"后期工具箱"。它按"你想做什么"来组织，每行都附上对应的工具示例与最关键的规则——规则栏里的内容，全部来自第五章失败模式的机制分析，是"用血换的"。

表 3-13　AI 视频后期工具箱（截至 2026 年 10 月）

你想做什么	典型做法	工具示例	关键规则 / 坑
局部重绘：换掉画面里一个物体、修一处细节	圈出（遮罩）要改的区域，输入替换描述，其余保持不动	即梦"精准局部编辑"、可灵视频编辑（主体与背景增删改）、万相视频编辑、Runway Inpainting 与 Aleph 2.0、Luma Ray3 Modify	遮罩外内容应当保持不变——生成后要专门检查"接缝帧"；视频编辑（V2V）是文字渲染出错的主要场景，画面里的文字应在这一步之前就规划为"后期叠加"
延长：把现有片段"接着拍"更长	以原片段尾帧为起点续写；或用"首尾帧模式"控制终点	可灵续拍、即梦"接着拍"、万相视频延长、Vidu（延长最长可至 5 分钟）、Premiere Generative Extend（每次最多 +2 秒）	延长次数越多衰减越重（图 3-16）；**保护最后几帧**、在提示词里请求"自然延续"而不是"新事件"；**不要在片段的第 1 帧处拼接**——接缝后第 1 帧常是错误帧，可靠运动通常要到第三个重叠帧左右才出现
补帧与慢动作：让运动更平滑，或做升格效果	在帧与帧之间生成中间帧；慢动作 = 插帧 + 放慢播放	剪辑软件的速度曲线与插帧功能；超分提升清晰度：MiniMax H3 超分 2K/4K、混元超分网络 1080p、PixVerse 视频超清	慢放会**放大**闪烁与形变——帧间差异被放大后更显眼；更稳的做法是**在生成端就写"慢速运动"**，把变速留给"锦上添花"而非"救场"
对口型与数字人：让角色对上语音	输入人像 + 语音，驱动口型与面部	可灵对口型与 AI Avatar 2.0、万相数字人（wan2.2-s2v）、PixVerse 平台级 Lip Sync、混元早期系列的对口型与动作驱动；国际侧 Veo / Kling 的原生"音画同出"	**声音与肖像都需要授权**——生成前留存授权凭证；数字人素材不得用于未获授权的真人（3.13 会展开）
字幕与画面文字：标题、字幕、LOGO	先生成"干净画面"，再在后期叠加所有文字	剪映的智能字幕等 AI 辅助功能（具体清单以官网为准）、Premiere / After Effects	这是失败模式"文字渲染"的标准解：**不要让模型渲染文字**——最强模型在评测中的整词错误率仍然可观；文字要求像素级一致，而扩散模型每帧都在重新采样
音效与配乐：给成片铺声音	音效先对齐动作，配乐后铺情绪	可灵 AI 视频音效、混元 Foley（视频音效生成）、Pika SFX、剪映曲库、Luma 集成的第三方音频	**音乐版权是红线**（第 1 章已展开 Suno/Udio 案例）：课堂作品优先用平台自带曲库并阅读条款
稳定化与降噪：救一救轻微闪烁与抖动	后处理滤镜：去闪烁、稳定、光流补帧	DaVinci Resolve 的 deflicker、Topaz Video AI 的稳定化	后处理是"降级补救"：能抹掉轻度闪烁，但**机制级的失败（身份漂移、物理错误）救不回来**——该重抽就重抽

关于延长的提示词，这里再给一张"可直接抄"的卡片——它体现了两个纪律（保护尾帧、请求延续而非新事件）：

✎ 提示词 • 延长请求（以 V1 尾帧为起点续拍）

从最后一帧自然延续：镜头继续缓慢推近，同一轮明月继续升离潮水，**[江面倒影]**保持被拉长的形态；薄雾流速不变，水面波纹幅度与前段一致。保持同一机位、同一焦段、同一月光光源；**不要引入新的人物或事件**，只做运动与时间的延续。无闪烁、无重影、无场景跳变。

用法 续拍 / 延长入口　起点 上一段尾帧　时长 +5 秒　禁止项 新事件、跳切

最后交代一下国产剪辑工具的位置：以**剪映**为代表的工具已经把"智能字幕""图文成片""自动配乐"之类的 AI 辅助做成了常规能力（具体功能以官网清单为准，2026 年 9 月其还发布了新的创作入口）——它们扮演的是流水线里"③ 编辑 → ④ 配乐"的宿主环境。对你而言最重要的判断是：**生成工具负责"造镜头"，剪辑工具负责"组织镜头"**，两边的技能都要练，但组织镜头的部分（节奏、景别衔接、轴线、蒙太奇）恰恰是第 1 章打下的底子——工具会换，听你指挥的还是那套镜头语言。

3.9 现实约束：一个镜头要"抽"几次才够

这是全章最"不技术"、却最重要的一节。如果你以为 AI 视频的工作方式是"写一句提示词 → 得到想要的镜头"，那么第一次实践课就会给你一个下马威。行业实践总结出的两条速记率，请现在记住：

*① 大约 **3 次生成** 才能得到 **1 个可用镜头**；　② 大约 **25%** 的生成片段会进入**最终剪辑**。*

两条速记率来自行业实测口径（不含 LoRA 等定向优化的普通工作流）。它们不是"上限"也不是"诅咒"，而是**规划预算与心理预期**的基准线：低于这个效率说明流程有问题，高于它值得写进笔记总结经验。

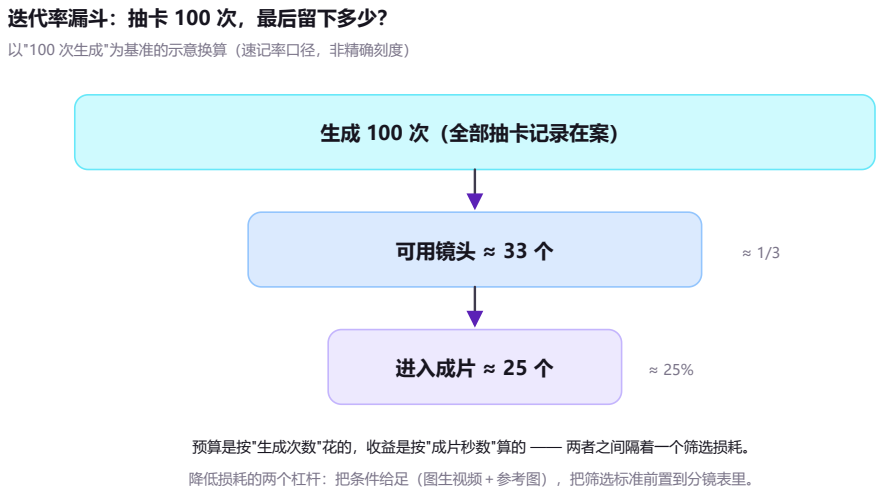


图 3-19　迭代率漏斗。"抽卡"不是失败的同义词，它是这套工作流的**固有成本结构**。专业团队与业余玩家的差别不在于"抽得少"，而在于**每次抽卡都记录参数**——让 3 次里至少有 1 次是"有方向的接近"，而不是三次纯随机。

这些比率折算成钱是什么概念？行业里有一个可引用的实例：**一个 70 秒、2 个角色的短片，在不用 LoRA 定向训练的条件下，生成花费约 750 美元。**换算一下：每 1 秒成片对应约 10 美元的生成预算（$750 ÷ 70 秒）——注意分母是**"最终成片秒数"**，不是"生成素材秒数"。这就是两条速记率的经济含义：你为"生成"付钱，却为"筛选后剩下的部分"拿到价值。另一组规模感更强的数据：一条 3 分钟、4 个角色的动画剧集，跑了 **164 个生成片段**——量级上，成片里的每一个"几秒钟"，背后都是几十次生成中的幸存者。

∫ 深入一格 · 预算怎么算：三个公式够用了 选学 · 成本与工程

- **素材量估算：**生成秒数 ≈ **成片镜头数 × 每镜尝试次数（≈3）× 单镜时长。**例：6 个镜头、每镜 5 秒、每镜抽 3 次 → 约 90 秒生成素材，对应成片约 30 秒（还要再按 25% 的留存率、按剪辑节奏缩一遍）。
- **生成预算估算：**生成预算 ≈ 生成秒数 × 平台"元/秒"单价。以万相 3.0 的 1080P 档（约 1.2 元/秒）为例，上面 90 秒素材 ≈ 108 元；换成 480P 档（约 0.3 元/秒）则 ≈ 27 元。**同一个项目，分辨率档位差 4 倍成本**——所以"低档抽卡、高档出片"是标准操作。
- **时间预算估算：**等待时间 ≈ 生成次数 ×（单次生成耗时 + 排队时长）。网页版工具的高峰排队不可控，**把"批量提交、离线等待"当作常态**，不要把渲染时间排进"交作业前的最后两小时"。

由此可以推出三条"降本纪律"，它们与 3.6 的参数语义互相呼应：

1. **探索用免费额度，交付用付费额度。**免费档通常带水印且不可商用，但它适合"试提示词、试参数、找气质"——把付费的每一次都花在已经验证过的组合上。
2. **低分辨率试参数，高分辨率出成片。**分辨率对成本的影响可达数倍，而"构图与运动是否成立"在低档就能判断——把分辨率留给最后一版。
3. **能用剪辑解决的，不要用生成解决。**需要 8 秒的运动，可以是两个 4 秒片段的"动接动"组接；需要"更多镜头"，可以把一个长镜头切片重组。剪辑比生成便宜一到两个数量级，这正是第 1 章镜头语言在这里的"变现"方式。

最后是心理账本。把这条流水线走到成片，比较常见的时间分配是：**生成（含排队与重抽）占大头，筛选与剪辑次之，参数调试也占相当比例。**所以本课程对作业的时间建议是**"提前三天开始生成"**——不是因为生成慢，而是因为你需要给"失败—诊断—重试"留出整整一轮的余量。下面 3.10 节的失败诊所，就是帮你把"重试"从盲目变成有方向的。

3.10 失败案例诊所：把"翻车"翻译成"处方"

本节按医院的方式组织：先量"症状"，再对"成因"，最后开"处方"。在这之前，先记住一条叙述纪律——**机制层引论文，实践层引厂商知识库**，两者分开表述。意思是：解释"为什么会这样"时要说清是研究结论（如"亚地球重力"、注意力时间对角线），给出"你该怎么做"时要说清是操作经验（如"保护最后几

帧"）。把经验做法说成机制、或把机制论文说成操作手册，都会误导后来的学习者——这也是你写实验报告时应遵守的同一把尺子。

先看全景。下表把九类高频失败模式压缩成一张速查表；请把"症状"列当作你的分诊台：**看到什么现象 → 去第三列找处方。**

表 3-14　九类失败模式总表（症状 → 成因 → 处方）

症状	成因（机制层）	处方（操作层）
闪烁：亮度脉冲、纹理抖动、边缘振动	注意力缺时序平滑归纳偏置（时间对角线碎片化）；训练-推理的噪声/SNR 分布差；"细节 vs 一致性"目标冲突；VAE 潜时间边界解码跳变	钉死单一光源；每镜只允许一个镜头运动；负面词 `no flicker, no warping`；缩短片段重抽；后期 deflicker 作为补救
形变 / 融化：物体边缘流动、结构糊掉	标准去噪目标缺乏几何一致性的显式激励；失去对先前帧的显式记忆；速度突变诱发形变；运动预算超支（多轴、俯仰拉伸细长特征）	锁身份（3–5 张多角度参考）；优先图生视频；每镜一个主要运动；慢动作；拆短片段；负面词 warping
物理错误：悬浮、下坠过慢、碰撞无反弹、液体像凝胶	"亚地球重力"：实测等效重力仅 0.38–2.27 m/s²（地球 9.81），且模型越大反而越慢；模型被训练成"看起来对"而非"行为正确"	拆 3–5 秒短片段；**不要安排精确的碰撞/液体/软体演出**（改用实拍或传统手段）；选题层面绕开弱项
面部崩坏：下颌线移位、鼻子变宽、面孔逐渐"溶解"	逐帧重新诠释角色、无身份锚点；模型优先保动作、其次保身份；文本与参考图竞争；单纯优化提示词只能解决约三成问题	参考包（正面、3/4 侧、侧面）同场次同光线；输入图规范（脸占画面高约 20–35%）；每镜重复 3–5 个核心特征；把"转身回头微笑"拆成两段
跨镜头不一致：换镜"变脸"、服装变化、画面里出现"双胞胎"	末帧链式生成会累积误差——到第 3、4 段就已不符；长时序间隔后记忆退化；工具之间没有持久的身份记忆	**改用参考生视频**：把完整上一段 + 原始参考一起喂入，每段重新锚定源参考；先做多角度角色设定表；每条提示都重复身份块
文字渲染错误：屏幕文字扭曲、笔画溶解、LOGO 变形	文字要求像素级一致，而扩散模型每帧都在概率采样——根本冲突；评测中最强模型的整词错误率仍达 0.250	**不要让 AI 渲染文字**：先生成干净画面，再后期叠加；负面词 `no text overlay, no watermark`
长片头尾漂移：纹理渐进退化、内容偏离上下文	因果/流式生成的误差累积 + 注意力感受野受限；自回归递归放大早期不一致；延长次数越多越差	缩短生成单元、靠剪辑；**避免在片段第 1 帧处拼接**（可靠运动约到第三个重叠帧才出现）；保护最后几帧、请求"自然延续"
手指 / 肢体重错、角色克隆	手的几何复杂度；生成之间不携带记忆，模型用"把同一角色画两次"解决歧义	构图规避：非必要时把手放到画外；慢动作；负面词 `no extra fingers`；固定角色设定表
运动幅度过大导致结构崩坏	运动预算超支；速度突变诱发形变；多角色身体接触最容易崩；POV 与过肩镜头是跨模型弱点	"缓慢、可控、连续、稳定"措辞；每镜一个主体、一个动作；空间约束写进每条提示；接受迭代成本（≈3 次生成 1 个可用镜头）

接下来是四间"诊室"，分别对应主线上真实遭遇过的四类问题。请重点观察每间诊室里"**诊断动作**"的写法——那是把症状对照机制的过程，也是你写实验报告"失败分析"一节时的模板。

🩺 **失败诊所 F3-1** 江面闪烁：每 6 帧亮一次的"鬼火"

症状： V1"月升·潮水"的第二版里，江面纹理从第 2 秒开始出现规律性明暗脉冲——肉眼看像有一层"鬼火"以每秒数次频率闪过；把时间线放大到逐帧，能看到水面亮度大约每 6 帧跳一次。构图、月亮的位置、雾的流动都很好，唯独光不对。

成因（机制层）： 对照 3.3 节的四种闪烁成因逐条排查：① 这条素材是 5 秒、单一机位、单一光源，运动预算没有超支，排除"冲突类"成因；② 亮度脉冲集中在**水面纹理**——细节密集、帧间像素变化剧烈的区域，正是"细节合成与一致性目标冲突"的高发区；③ 脉冲呈"每隔几帧"的规律性，与"分块生成时潜时间边界的解码跳变"特征吻合。综合判断：**主因是②+③的叠加**，运动的复杂程度只是放大器。

处方（操作层）： ① 提示词"稳定优先"：明确"水面只做轻微波纹起伏"、钉死单一光源（月光）；② 负面词补齐 `no flicker, no warping`；③ 把片段从 5 秒缩到 4 秒重抽（缩短片长直接减少潜边界数量）；④ 后期用 DaVinci Resolve 的 deflicker 作为补救——但只用在"其他方面近乎完美"的素材上，不要拿它当常规流程。

复诊： 重抽结果的验收方式：把画面缩到 50%、逐帧检查水面区域——闪烁在缩小视图上依然可见才算"未通过"。最终第 3 次生成的版本通过复诊，进入成片。

🩺 **失败诊所 F3-2** 小船"融化"：舟形蠕动与人影双头

症状：V5"孤舟"第一版：船身中段像"液态金属"一样缓慢蠕动，船头与船尾对不上；划桨人影附近出现了第二个头部轮廓；这三种现象在不同帧里交替出现，没有一帧完全干净。

成因（机制层）：① **运动预算超支**——这一镜同时叠加了划桨动作、船体漂流、镜头横移、水面波纹四个运动源；② **速度突变**——原始提示词写了"由慢突然加快"，正是论文里明确会"诱发形变伪影"的写法；③ **身份未锚定**——这一镜当时是纯文生视频，船与人没有任何参考图锚点，模型只能逐帧"重新想象"船的形状（"失去对先前帧的显式记忆"）。

处方（操作层）：① **拆镜**：把"划桨特写"与"孤舟横越"拆成 V5a / V5b 两个镜头，让每镜只有一个主要运动；② **锚定**：V5b 改为图生视频，首帧直接用 K5 分镜图，船形从此有据可依；③ **降运动**：运镜改"固定"（横越感由主体漂移产生，而不是镜头横移——这同时解决了"运动叠加"）；提示词删掉"突然"，只保留"缓慢、连续"；④ 负面词加 `no warping, no duplicate characters`。

复诊：重拍后逐帧检查"船首—船身—船尾"三点轮廓的连续性，以及人影数量；任何一帧出现第二个头轮廓即打回。第 2 次生成即通过。

🩺 **失败诊所 F3-3** 花瓣像在糖浆里下坠

症状：V3"花林"里安排了一片飘落的花瓣作为前景点缀；结果花瓣的下落速度明显偏慢，轨迹"飘飘悠悠"，像在糖浆里下沉；后续一版试图用"加速下落"提示词补救，反而出现花瓣边缘拉伸。

成因（机制层）：这正是 3.3 节讲过的"亚地球重力"现象的直接表现——实测中模型的等效重力只有 0.38–2.27 m/s²，远低于地球的 9.81 m/s²，且大模型并不比小模型更"物理"；模型被训练成"看起来对"而不是"行为正确"，像素层面的似然目标不奖励物理守恒。另外，"加速下落"属于速度突变指令，触发形变伪影——**用错误的处方治症状，只会引入新症状。**

处方（操作层）：① 不要在低重力世界里硬演重力戏——把花瓣从"前景主体"降级为"背景氛围元素"（缩小、虚化、放在画面边缘），观众的注意力不再落在那条轨迹上；② 若必须有下落动作，拆成 3 秒以内的短片段，并接受"轻度不真实"；③ 记住这条边界：**精确的碰撞、液体、软体演出不应该安排给视频模型**——这属于实拍或传统特效的守备区；④ 想从根上改善物理表现的路径（学术侧）：用少量单球下落片段训练轻量 LoRA，可把等效重力从约 1.81 提升到 6.43 m/s²（约地球的 65%）——这说明"物理短板"是可以被定向训练的，但那是进阶课题。

🩺 **失败诊所 F3-4** 匾额上的字，每帧都在"呼吸"

症状：V4"楼台"的中景里，匾额上的两个字在 5 秒内笔画间歇性溶解又重组——暂停看每一帧都像"差一口气就写对了"，播放起来像"字在呼吸"。这一镜的构图与栏杆线条都很稳，唯独文字无法接受。

成因（机制层）：文字要求**像素级一致性**——同一个字在 120 帧里必须笔笔相同；而扩散模型是逐帧概率采样，"每一帧都是重新采样"。评测数据（VTR-Bench，300 条提示、11 个模型）显示最强模型的整词错误率仍达 0.250；研究还发现"内容识别"与"字形级正确"是分离的能力，而视频编辑（V2V）是文字的主要瓶颈。需要说明：**"汉字是否比拉丁字母更难"目前尚无专门的实证研究可引用**，本教材只作方向性判断——笔画密度更高的字符，对逐帧一致性只会更不利。（存疑事项，勿夸大）

处方（操作层）：① 标准解：**不要让模型渲染文字**——把匾额生成成"干净的、不可辨认的装饰笔画"，成片后在后期用剪辑软件叠加清晰文字（也更方便替换与校对）；② 负面词加 `no text overlay, no watermark`；③ 若剧本坚持"画面内自然出现文字"，换成远景或大俯拍——让文字小到不构成"读认对象"，模型只需画对"有一块匾"，不需要画对"匾上是什么字"；④ 小组纪律升级：复查第 2 章分镜表，凡含文字元素的镜头在生成前就标注"后期叠加"。

🩺 **案例 3-B** § V5 孤舟的完整重拍：从"翻车"到"入库"

背景：主线《春江花月夜》推进到第 5 个镜头"孤舟"（对应 K5 分镜图：远景平视，一叶小舟自右向左横越江面）。这一镜是全片运动最复杂的镜头，也是小舟最"志在必得"的一镜——她原以为前面四个镜头都过了，这一镜只是重复劳动。

第一版（失败）：纯文生视频、5 秒、运动幅度中档、运镜"镜头横移 + 轻微升镜"，提示词写了"划桨加速、由慢突然变快"。结果：船形蠕动、人影双头、水面闪烁——三种症状同时出现（详见 F3-2），判定不可用。她又原样重抽了两次，结果只是"换了一批毛病"——没有诊断的重抽，只是在重复付费随机的过程。

诊断（对照 3.10 诊所表）：她把三版结果并排，逐条对照表 3-14 的"症状 → 成因"：运动预算超支（四个运动源叠加）+ 速度突变（"突然"二字）+ 身份未锚定（无参考图）。结论：这不是"运气不好"，而是**三张处方同时缺席。**

重参数（三张处方）：① 拆镜成 V5a"划桨特写"（4 秒、固定机位、只保留桨的运动）与 V5b"孤舟横越"（4 秒）；② V5b 改图生视频，首帧 = K5 分镜图，另附 2 张船形参考；③ 运镜改"固定"，提示词改为"孤舟自右向左缓慢漂移，速度恒定"，负面词补齐。她还做了第四件事——**把这次实验的完整记录写进了表格**（平台 + 版本 + 参数 + 症状 + 处方），因为"下一个学弟学妹会需要它"。

结果与复盘：V5a 第 1 次生成通过；V5b 第 2 次生成通过，仅需后期轻微去闪烁。对比第一版的 3 次全废，重参数后的两次里投产两镜。复盘一句话：**失败时不要重抽，要换条件**——重抽是买彩票，换条件是做实验。这条复盘后来成了小组的工作口令，也写进了他们的答辩幻灯。

把四间诊室的处方放在一起看，会发现高频出现的动作其实只有三个：**给参考图（锚定身份）、拆镜头（降运动预算）、缩片段（缩短误差累积）**。它们共同指向一个事实：AI 视频的失败，多数不是"提示词不够华丽"造成的，而是"条件给得不够"造成的。这也是本教材把条件控制（3.5）排在失败诊所（3.10）前面的原因——先学开药，再学看病。

最后，请把失败当作**数据**而不是挫折：每次失败都按"症状—成因—处方"三栏记录下来（3.12 节提供了记录表模板），一学期的积累会让你手里长出一张只属于你自己的"故障地图"——这张地图的价值，超过任何教程。

3.11 关键概念表：把这一章拧成一把钥匙

下表是本章全部术语的"最小集"：每个词一句话定义，外加一条最容易搞错的提醒。做作业卡壳时先回来查它；答辩被问到术语时也用它自检——能用自己的话把第三列说清楚，才算真的懂了。

表 3-15　本章关键概念与易错点

术语	一句话定义	最容易搞错的点
帧 / 帧率（fps）	视频的最小单位是单帧图像；帧率是每秒帧数	"5 秒视频 = 5 张图"是错的：24fps 的 5 秒是 120 帧——这是算力差的根源
时空数据立方体	视频是"帧数 × 高 × 宽"的三维数据块（T×H×W）	多出来的时间维不是"多存几张图"，它让注意力的计算量按平方增长
时空 patch / 词元	把立方体切块、编码后交给 Transformer 的输入单元	"时空 patch"是**输入表示**，不等于"时空注意力分解"——Sora 恰恰不属于分解式路线
潜空间 / 潜扩散	先把视频压缩到低维表示，在潜空间里做扩散去噪，再解码回像素	压缩比越高不等于越快——编解码常是隐藏瓶颈（3.6 的反例）
3D 因果 VAE	在空间与时间维都做压缩的自编码器（典型约 4×8×8）；"因果" = 只看过去帧	SVD **没有** 3D VAE——它的时间层在 UNet 里去，其 VAE 沿用图像版；这一说法在教程里常被传错
时间混合层	插进 2D 生成器、专门建模帧间关系的层	只加时间层、不做时空联合建模时，长程一致性有限、易闪烁
全时空注意力 / 分解式注意力	所有词元两两互看（$O((THW)^2)$） / 帧内空间 + 同位置跨帧（省但一致性略逊）	这是"成本 vs 质量"的取舍，不是"新旧"的替代关系
DiT（扩散 Transformer）	用 Transformer 替代 U-Net 做扩散主干的架构	"DiT"说的是主干结构，与"如何处理时间"是两个独立维度
注意力平方增长	词元数 N 翻倍，注意力计算量变 4 倍	分辨率从 1080P 到 4K（像素 4 倍），注意力约 16 倍
空间一致性 / 时间一致性	身份·布局·风格·光照一致 / 运动平滑·状态转移合法·低闪烁·动作忠实	"单帧好看"不等于"动态连贯"——审视素材时要分开打分
闪烁（flicker）	亮度脉冲、纹理抖动、边缘振动的**高频**不连贯	它不是单一起因：注意力缺平滑偏置、训练-推理 SNR 差、细节-一致性冲突、VAE 边界都可能是主因
漂移（drift）	内容随时间**低频**地偏离全局上下文（越走越远）	与闪烁要分开诊断：闪烁在"帧间"，漂移在"片段间"
首尾帧插值	给首帧与尾帧，让模型生成中间过程	两帧差异过大时模型会自行编造过渡；部分接口固定时长（如 5 秒不可改）
参考图 / 主体一致性	用若干张参考图把"是谁"锚定下来	参考图与参考视频的配额会互相挤占；用量与上限以平台文档为准
运动控制	用轨迹、相机位姿控制运动（研究侧）；网页版多以提示词或参考视频复刻实现	轨迹近邻或交叠时容易"串轨"；别把"提示词运镜"当成位姿级控制
音画联合生成	图像潜与音频潜在同一次去噪中生成，同步是"学"出来的	与"先生成再配音"的后配音路线是两条路，别混为一谈
"世界模拟器"叙事	Sora 报告提出的长期愿景：视频生成走向物理世界模拟	它与其反证并存：视觉真实感与物理理解无关（Physics-IQ 结论）
"亚地球重力"	实测 AI 视频中物体下落的等效重力仅 0.38–2.27 m/s^2（地球 9.81）	模型更大并不更"物理"——大模型反而更慢；这不是帧率伪影
种子（seed）	控制随机性的数值：同种子 + 同输入 = 同输出	只在**同模型同版本同参数同平台**内成立；网页版多不暴露种子；模型升级即可能失效
运动幅度	控制画面"动多少"的参数	量纲跨平台不可比（0–10、0–1、联动参数并存）；人像比风景更怕高幅度
单次时长 / 累计上限	一次生成的最长长度 / 拼接与延长后的总长度	不是一个量纲：AI 视频的时长是"拼"出来的，不是"生成"出来的
迭代率	≈3 次生成换 1 个可用镜头；≈25% 的生成片段进入成片	这是成本结构，不是"水平不行"；写预算与排期必须用它

术语	一句话定义	最容易搞错的点
显式 / 隐式标识	用户可明显感知的标识 / 藏在文件元数据里的技术标识	两者都不得恶意删除或篡改；导出后别用工具"清理元数据"（3.13 展开）

3.12 动手实践：把分镜"动"起来

本周实践共 2 学时（90 分钟课堂） + 课后主线任务。课堂三个练习各有侧重：**实践一**建立"条件决定命运"的直觉；**实践二**练参数对照实验的方法论；**实践三**练镜头语言的"提示词翻译"。课后把主线推进到"4–6 段镜头视频"，并为其中至少一次失败留下完整档案。开始之前先做两件准备：**① 选一个国内网页工具固定下来**（可灵 / 即梦 / 海螺 / Vidu / 万相均可，选能看历史记录的那个）；**② 准备好第 2 章交付的分镜图（K1–K6）与角色设定图**——它们是本次实践的全部"弹药"。

🛠 实践一：同一分镜的两条路——文生视频 vs 图生视频（30 分钟 • 课堂）

基础

任选一张分镜图（建议 K1"月升·潮水"，它的运动最简单），用同一段提示词跑两组：**A 组纯文生视频**（只给文字）与 **B 组图生视频**（把分镜图当首帧）。两组各生成 **3 次**；除"是否给首帧"外，一切参数保持相同（时长、比例、分辨率、运动幅度、种子策略一致）。

按三个维度打分并写理由：① 与分镜表的一致性（构图、色调、意象）；② 运动的合理性（方向、速度、有没有物理错）；③ 瑕疵统计（闪烁/形变/多余肢体各出现几次）。

提交物：6 条视频 + 一张"三列打分表"。结论请回答一个问题：**图生视频把失败概率从哪一类"降级"到了哪一类？它消灭了失败吗？**（参考案例 3-A 的复盘。）

进阶

做一个"提示词杂交"实验：把 A 组里运动气质最好的那条视频的提示词片段（描述"怎么动"的那句）抄进 B 组的提示词，重生成 2 次。观察命中率是否提高，并写一句心得：**文生视频有时是"素材库"，图生视频是"装配线"。**

🛠 实践二：参数对照实验——运动幅度与时长（40 分钟 • 课堂）

进阶

选择"人像中景"与"风景远景"各一个镜头（建议 K2 与 K1），按下表跑三组实验。**铁律：一次只改一个变量，其余全部锁死并写进记录**；每组至少生成 2 次，以区分"参数效果"与"随机差异"。

表 3-16　参数对照实验设计（一次只改一个变量）

实验	变量（只改这一个）	三组取值	固定不变项	观察点	通过标准（开始前写死）
实验 A • 运动幅度	运动幅度 / 运动强度档位	低档 ｜ 平台默认 ｜ 高档（按平台量程折算，如 0–10 量程取 2 / 5 / 8）	提示词、时长、比例、分辨率、是否图生、参考图	主体形变率、背景稳定性、动作完成度、画面"活"的程度；人像与风景分开记	能给出两组结论：人像的"形变开始明显"的档位 vs 风景的"还能更高"的档位——并解释为什么人像更脆弱
实验 B • 时长	单次时长	4 秒 ｜ 5 秒 ｜ 8 秒	提示词、运动幅度（取实验 A 的中间档）、分辨率、是否图生	末段漂移、闪烁出现的位置（第几秒）、单帧锐度、等待时间	能指出"瑕疵从第几秒开始出现"；能复述"4–10 秒是基本生成单元、长片靠拼接"
实验 C • 落幅	提示词是否写"落幅"	只写运镜名 ｜ 运镜 + 速度 ｜ 运镜 + 速度 + 落幅	其余提示词、参考图、时长、运动幅度	镜头结束时停在哪、能不能接下一镜、构图是否回到可剪辑状态	能当场向同学展示"没写落幅"的版本为什么不能用于剪辑——用画面证明纪律的必要性

挑战

跨平台量纲测试（选做）。把同一个镜头素材交给第二个平台，用"同一个意图"（而非同一个数值）重设运动幅度与时长，比较结果差异。结论写成一段话：**参数值只在各自平台内可比；跨平台要用"意图"翻译，不能用"数值"搬运。**

🛠 实践三：运镜指令的"翻译"练习（20 分钟 • 课堂）

进阶

从图 3-13 的九种运镜里任选 **3 种**（其中必须包含"环绕"），把每种写成一条完整提示词——必须包含：主体描述、运动方式、运动幅度、运镜类型、景别与角度、**落幅**。然后各生成 1 次。

自检三个问题：① 视频结束时镜头停在哪里？和你的落幅写的一致吗？② 有没有出现"多轴叠加"（例如环绕中还自带推拉）？③ 如果重写一次，你会删掉哪个词、补上哪个词？

提交物：3 条提示词 + 3 条视频 + 每条的"改一词"结论。这条练习同时是 30 度原则与景别级差的准备——下一镜怎么接，取决于这一镜"停在哪里"。

🛠 实践四（主线）：把 4–6 段镜头变成一部小片（课后 • 2–3 天）

挑战

任务：用第 2 章的 K1–K6 分镜图，按 3.8 节的流水线完成一部 **20–40 秒**的短片（成片 4–6 个镜头）：生成 → 筛选 → 编辑 → 配乐 → 导出。全程遵守三条纪律：**每镜全参数留档**（平台 + 版本 + 种子 + 所有非默认参数）；**筛选标准前置**（在生成前就把"什么算可用"写死）；**失败必写档案**（用下表记录，至少 3 条真实记录，其中至少 1 条做到"失败 → 诊断 → 重参数 → 成功"的完整闭环，参考案例 3-B）。

剪辑要求：镜头组接要遵守第 1 章的规则——景别级差两档以上、动接动方向一致、写清每镜落幅；片尾按 3.13 节完成 AI 标识。

表 3-17　失败案例分析与改进方案记录表（模板 + 示例）

症状（可观察描述）	镜次 / 版本	对照失败模式（表 3-14）	成因假设（机制层）	处方动作（操作层，已执行）	复诊结果	结论与下次预防
第 2 秒起江面纹理规律性明暗脉冲（约每 6 帧一次）	V1 · 第 2 版	#1 闪烁	细节-一致性冲突 + 潜时间边界解码跳变（运动本身不复杂）	钉死单一光源；负面词 no flicker；5 秒缩 4 秒重抽	第 3 次通过复诊（50% 视图逐帧）	水面/雨雪等高频纹理镜头：默认缩短到 4 秒起步
船身中段蠕动、人影出现第二个头轮廓	V5 · 第 1 版	#2 形变 / #9 运动超支	四轴运动叠加 + 速度突变 + 无身份锚点	拆镜（V5a/V5b）；改图生视频；运镜改固定；删"突然"	V5a 1 次通过；V5b 第 2 次通过	结构复杂主体（船/车/机械）：一律先拆运动、再谈运镜
（示例结束，下面是你要填写的）	—	—	—	—	—	—

提交物清单

① 成片视频文件（命名：`组名_春江_终版.mp4`）；② 参数与素材记录表（镜头号 / 平台 / 版本 / 种子 / 关键参数 / 尝试次数）；③ 填好的失败分析记录表（表 3-17，≥3 条）；④ 一份 200 字以内的方法复盘：本次实践里"最省钱的一次决策"是什么？

3.13 真实与标识：让 AI 作品"说真话"

学会生成之后，你必须面对一个比任何参数都重的题目：**当你造出的画面与真实几乎无法区分时，你要如何让别人知道"这是 AI 做的"？**这不是道德说教课，而是一门有法可依、有案可查、有作业标准要执行的"手艺"。原因很简单：视频的"可信度"高于任何其他媒介——人们本能地相信"眼见为实"。当"眼"可以被合成，信任就成了需要制度与习惯共同维护的稀缺品。

先看反面教材。过去几年里，利用"高可信媒介 + 生成技术"的欺骗已经形成了完整产业链，而且无一例外地指向最容易相信"眼见为实"的人群：

- **AI"靳东"诈骗老人案：**假冒明星粉丝团诈骗 30 余万元——**2024 年 2 月上海静安区法院一审判决 8 名被告，主犯有期徒刑 3 年、罚金 5 万元**；此后仍出现"江西老人欲贷款 200 万元给'男友靳东'"的案例；2025 年两会期间靳东本人回应称"性质非常恶劣"，并建议对 AI 换脸立法。
- **张文宏"AI 带货"：**AI 合成视频卖蛋白棒，显示**售出 1,200 余件**；2024 年 12 月上海市消保委认定其**构成对消费者欺诈（可主张退一赔三）**；微信平台专项治理处置内容 532 条、关闭账号 209 个。
- **奥运冠军 AI 声音带货：**2025 年 8 月 18 日央视新闻曝光：AI 仿冒全红婵声音卖土鸡蛋，单条视频点赞 1.1 万，链接**售出 4.7 万件**；卖家声称"一分钟素材即可克隆任何人声音"。
- **司法给出的价码：**杭州"AI 换脸"民事公益诉讼案 2023 年判赔 **6 万元**并公开道歉；全国首例 AI 生成声音人格权案（北京互联网法院，2024 年 4 月 23 日）判赔 **25 万元**；公安部 2023 年 8 月的 AI 换脸专项会战通报侦破 79 起、抓获 515 人。

写这段时教材编辑部还做了一个"反面示范"——网络上流传过一个数据："AI 换脸诈骗同比增 300%"。它看起来很适合放进教材，但我们核查后发现**该说法仅见于内容农场**，无法溯源，因此不予采用。这本身就是你在这门课里要养成的习惯：转述任何"惊人数字"之前，先问它从哪里来——生成式 AI 让造假成本趋近于零，也让"核查"成了创作者的基本功。

面对这些风险，中国的选择是"**办法 + 强标**"的治理模式：2025 年 9 月 1 日起施行的《人工智能生成合成内容标识办法》（四部门联合发布），配合当日实施的强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》（GB 45438-2025），搭建起两层标识体系：

- **显式标识：**在内容或界面中添加"可被用户明显感知"的标识。对**视频**，要求在**起始画面和播放周边适当位置**添加显著提示标识，可在末尾和中间适当位置添加；对音频，要求在起始、末尾或中间添加**语音提示或音频节奏提示。**提供下载、复制、导出功能时，要确保文件里含着合规的显式标识。
- **隐式标识：**在文件**数据**中添加"不易被用户明显感知"的标识——**元数据**中写入生成合成内容属性、服务提供者名称或编码、内容编号等；并鼓励添加数字水印。平台在传播环节还要核验隐式标识、补加"传播要素信息"。

- **红线**：任何组织和个人**不得恶意删除、篡改、伪造、隐匿标识**，也不得为此类行为提供工具或服务。这直接标明了两件"不能做"的事：给 AI 视频"洗元数据"，以及裁掉平台水印假装没生成过。

法规不只是纸面。把"执行情况"也讲清楚，才是这门课该有的实证态度：网信部门 **2025 年 11 月**集中查处过一批未有效落实标识规定的 App（约谈、责令限期改正、下架下线）；据新华网报道，**2026 年 4 月 28 日**，网信部门发现"剪映""猫箱"App 及"即梦 AI"网站未有效落实标识要求，指导属地网信办采取**约谈、责令改正、警告、从严处理责任人**等处置处罚——被点名的正是本章工具表里最主流的几个国产工具。此外，**2026 年 3 月起**，抖音、快手、腾讯、小红书、哔哩哔哩、微博等 **12 家平台**先行先试短视频标注，提供"含 AI 生成内容"等 **6 类必选标签**，标注成为发布的必经环节；**2026 年 1—5 月**，清理虚假摆拍等违规短视频 **52 万余个**、严惩违规账号 **6.8 万余个**；为期 4 个月的"清朗·整治 AI 应用乱象"专项行动第一阶段就处置违规产品 1.4 万余款。

同时，也必须把"执行还在路上"的另一面写出来，这才是完整的现实：媒体实测发现——2025 年 4 月记者测的 25 家平台里仅 9 家设有 AI 声明入口；**2026 年 4 月，仍有不少短剧中的 AI 片段未带标识，无标识的 AI 视频"3 秒即过审"**。一项覆盖 40 家平台的研究还发现：仅有 14 家强制披露，其中 6 家只对"真实感内容"有要求，且自动标注普遍没有申诉渠道。有法可依 ≠ 执行到位——所以你不能把"平台会替我负责"当作策略；责任是创作者的。顺带一个反向的提醒：标识也有"错标"的风险——北京互联网法院 2025 年 7 月的首例"平台判定用户内容系 AI 生成"案中，用户被算法判定"含 AI 内容未标识"而遭禁言，法院判**平台未尽合理适度说明义务、承担违约责任**。标识是一件严肃的事，两头都有责任。

把视角抬高一格：中国的"办法＋强标"不是孤例。欧盟《人工智能法案》第 50 条的透明度义务自 **2026 年 8 月 2 日**起适用：生成式 AI 输出须有**机器可读标记（至少两种技术，如签名元数据＋C2PA 式水印）**，深度伪造须显著披露（艺术创作可"有限披露"），罚则最高可达 **1,500 万欧元或全球营业额的 3%**（既有系统宽限至 2026 年 12 月 2 日）。平台侧的技术标识也在并行推进：Google 的 Veo 视频全部带 **SynthID** 水印；Adobe Firefly 生成的素材带 **Content Credentials**；TikTok 依据 C2PA 检测自动加"AI-generated content"徽章且不可移除。**三法域、两条技术路线（法规标识与技术水印）指向同一个方向：AI 内容的"可识别性"正在成为基础设施。**

> 🇨🇳 **课程思政 · 真实是创作的底线，标识是底线的告示牌**
>
> 《春江花月夜》里有一句话，很适合放在这一节的结尾："人生代代无穷已，江月年年望相似。"诗人用"不变的江月"照见"流转的人生"——文学的力量恰恰来自**把真实的世界转化为可信的表达**。生成式 AI 给了我们前所未有的表达自由，但它同时提出了一个古老的问题：**当你可以造出任何"真实"，你还愿不愿意对自己的画面负责？**
>
> 本课程的立场是三条职业信仰，每一条都对应可执行的动作：**第一，不拿合成冒充真实**——AI 视频不用于伪造新闻、伪造他人言行、伪造证据；虚构作品要标注"演绎"，涉及历史与新闻的内容不做"看起来像实拍"的模糊处理。**第二，尊重每一个真人的形象与声音**——肖像、声音、姓名都需要授权；数字人素材要能追溯到授权凭证；绝不"拿明星的脸练习"。**第三，把标识当成作品的一部分**——就像片尾字幕一样，AI 标识不是"缺陷声明"，而是创作者的诚实签名。观众可以原谅"这是 AI 做的"，但很难原谅"你假装它不是"。

落到操作，发布前只跑一遍下面这张流程图与检查清单即可。

先说明边界：「素材来源、真人权利、显式与隐式标识的最低技术要求」这套通用清单，第 2 章的表 2-23 已经逐条给出，附录 E 也有一份可打印版——**本节不重复那部分**，只讲**视频特有**的三件事：标识要出现在哪里、音频轨怎么办、以及成片发布与图像投稿在要求上的差别。

而下面这张流程，把法规条文翻译成"你手上的四个动作"：**判断、加显式、保隐式、留平台声明。**

图 3-20　发布前标识操作流程。五个步骤里，②③ 是法规动作（不可省），④⑤ 是职业动作（保护你自己）。注意它排在流水线的"导出"环节之后——**标识不是额外工作，而是导出的一部分**。

表 3-18　发布前标识检查清单（课程标准）

检查项	依据要点	本课程作业标准	常见错误
视频显式标识	视频需在起始画面和播放周边适当位置加显著提示	起始画面一角固定"AI 生成"提示 + 发布简介首行注明"含 AI 生成内容"	只字不提；以为"平台会自动加"
音频显式标识	音频应在起始、末尾或中间加语音或节奏提示	片尾一行语音："本片内容由 AI 辅助生成"	以为"画面有字就够"——音频内容也有独立要求
隐式标识完整	元数据应含生成合成属性信息；不得恶意删除篡改	导出后不做"元数据清理"；保留平台生成的原始下载文件	用"瘦身工具"处理文件，把标识一起洗掉
真人权利	肖像、声音、姓名受法律保护，伪造他人言行有明确判例与罚则	每段真人素材都有授权记录；不使用公众人物的脸与声音做练习	拿明星形象"练习"；数字人素材没有可追溯来源
内容真实性	AI 合成不得用于欺骗性传播	虚构内容标注"演绎"；历史、新闻性内容不使用"似实拍"的合成	把 AI 视频当"实拍花絮"发布
平台声明	12 家平台 2026-03 起试行"6 类必选标签"，标注是必经环节	逐个勾选声明项并留存截图	漏勾"含 AI 生成内容"，或发布后不复核
素材版权	音乐、图片、字体同样受版权约束（参照第 1 章音频版权处）	配乐优先平台曲库；每份素材的来源与授权入档	直接搬运网络音乐与字体

最后把"学生该怎么给自己的作业做标识"总结成三句话，请直接抄进你们小组的交接文档：① 作品里含 AI 生成的画面或声音，就在**视频起始画面加角标、简介注明"含 AI 生成内容"**；② 发布时**勾选平台的 AI 声明**，并把"带标识的成片 + 声明截图 + 参数记录"一起归档；③ 永远不为了使作品"看起来更像实拍"而删除、修改标识——那不是聪明的运营技巧，而是法规明令禁止的行为。会标识的创作者，才是可以长期合作、可以商业交付的创作者——这门课教你生成，也教你被信任。

◆ 把这一章压缩成七句话

1. 视频 = 图像 + 时间维：素材是"帧数 × 高 × 宽"的时空数据立方体；词元数随时长与分辨率线性增长，但注意力的计算量按**平方**增长——"多一维"就是"贵一档"。
2. 三条技术路线是"演化叠加"而非"三选一"：潜扩散时间层 → 时空注意力（全时空 / 分解式）→ DiT；当代主流是「3D 因果 VAE + 全时空注意力 DiT」。两处常见错说必须修正：**SVD 没有 3D VAE**；**时空 patch ≠ 注意力分解。**
3. 时间一致性要分两层看：空间一致性（身份、布局、风格）与时间一致性（运动平滑、低闪烁）；典型崩坏是**闪烁**（高频抖动）与**漂移**（低频偏离），根源是误差累积与注意力缺少时序平滑的归纳偏置。
4. "世界模拟器"愿景与实测反证并存：VideoPhy 显示最强模型仅 39.6% 同时满足提示与物理；Physics-IQ 的结论是"视觉真实感与物理理解无关"；"亚地球重力"实测 0.38–2.27 m/s²——**看起来真 ≠ 懂物理。**
5. 条件控制是质量的主开关：**先锚后动**——图生视频、参考图、首尾帧把"是什么"锚住，提示词再讲"怎么动"；运镜写作守"一镜一个主要运动 + 落幅 + 主体关系"三条纪律。
6. 参数跨平台量纲不可比（运动幅度有 0–10、0–1 等多种口径）；**时长是"拼"出来的**（4–10 秒为基本生成单元）；种子只在同模型同版本内可复现；分辨率"贵不贵"取决于瓶颈在扩散还是编解码。
7. 现实约束记两条速记率：**≈3 次生成换 1 个可用镜头、≈25% 生成片段进入成片**；失败时"换条件"而不是"重抽"；发布前完成 AI 标识（显式 + 隐式）——合规是创作流程的一部分，不是附加题。

◆ 自测

❓ 自测 1：为什么说视频比图像"贵一档"？请用"时空数据立方体"与"注意力的平方增长"两个概念解释。

答案：图像是"高 × 宽"的二维数据，视频是"帧数 × 高 × 宽"的三维数据块（时空数据立方体）。模型把数据切块编码成词元后，词元数与"时长 × 分辨率 × 帧率"成正比——这是**线性**增长；但注意力要求每个词元与所有词元交互，计算量约等于词元数的**平方**。于是：时长翻倍（词元翻倍）→ 计算量约 4 倍；分辨率从 1080P 到 4K（像素 4 倍）→ 注意力约 16 倍。这就是"多一维等于贵一档"的数学来源，也是"AI 视频的时长是拼出来的，不是生成出来的"的底层原因。

❓ 自测 2：同学在报告里写了两句话——"SVD 用了 3D VAE""Sora 的核心是分解式时空注意力"。请指出错误并改正。

答案：两句都是流行误说。① SVD 的时间建模发生在 **UNet 去噪器**里（插入时间混合层并联合微调），它的 VAE 沿用图像版（只有空间压缩、没有时间压缩）；带时间维压缩的 3D（因果）VAE 是 CogVideoX、Wan、混元等后续模型的做法。正确说法是"SVD 时代 = 图像 VAE + UNet 时间层"。② "时空 patch"是 Sora 的**输入表示 / 词元化方式**，不等于**注意力如何连接**；把二者混为一谈会错误归类 Sora 的技术路线。词元化与注意力结构是两个独立的维度。

❓ 自测 3：闪烁与漂移怎么区分？各给一条判断依据与一条处方。

答案：闪烁是**帧与帧之间**的高频不连贯——亮度脉冲、纹理抖动、边缘振动，逐帧检查最明显；漂移是**一段时间内**的低频偏离——身份、色调、场景元素越生成越"不像开头"，通常要对比首尾才看得出来。判断依据：放大时间线逐帧看（闪烁） / 把首尾帧并排看（漂移）。处方示例：闪烁——钉死单一光源、负面词 `no flicker`、缩短片段重抽，后期 deflicker 补救；漂移——缩短生成单元、靠剪辑而不是靠延长、用参考图/参考生视频把每一段"重新锚定"到源参考。

❓ 自测 4：有人看完一段"物理无懈可击"的 AI 视频后断言"模型已经理解了物理"。请用两条硬证据反驳，并说明这对应届时的什么实践决策。

答案：证据一：VideoPhy 评测中，当时最强模型也只在 39.6% 的实例中同时满足提示与物理规律；证据二：Physics-IQ 用 396 段真实物理实验视频评测，结论是"物理理解严重受限，且与视觉真实感无关"；补充证据：单球下落实测的等效重力仅 0.38–2.27 m/s²，且模型越大反而越慢——说明不是帧率伪影，而是系统性偏差。实践决策：不要让视频模型承担**精确的碰撞、液体、软体、多物体交互**演出，把这类镜头交给实拍或传统特效；在剧本阶段就"绕开弱项"，而不是在生成阶段反复硬碰。

❓ 自测 5：小组想在两个平台上比较"运动强度"效果，计划把 A 平台的 6 和 B 平台的 0.6 都设为"中档"再对比。这个设计哪里有问题？正确的做法是什么？

答案：问题在于"量纲不可比"：A 平台可能是 0–10 的量程，B 平台是 0–1 的量程，甚至还有平台让运动强度与其他参数联动——"6"与"0.6"在各自平台内的语义可能完全不同，直接对照等于用两把不同的尺子量同一条边。正确做法：① 先查各自官方文档确认量程与默认值；② 用"意图翻译"而非"数值搬运"——例如都以"人像轻微活动（眨眼、微微转头）"为同一意图，请各自平台折算到对应档位；③ 每档至少生成 2 次以区分参数效果与随机差异；④ 把量程、默认值、取值一并写进记录（附平台与版本）。

❓ 自测 6：把一部 AI 生成短片发布到平台前，至少要做哪三件事？为什么"删掉文件元数据"是错的？

答案：三件事：① 加**显式标识**——视频在起始画面与播放周边放显著提示（如"AI 生成"角标），音频加语音提示；② 保**隐式标识**——保留文件元数据与平台水印，发布时完成平台的 AI 声明/必选标签；③ **归档**——把标识位置、声明截图、生成参数与授权凭证存入创作档案。删除元数据之所以是错的：它直接命中法规明令禁止的行为——"任何组织和个人不得恶意删除、篡改、伪造、隐匿标识"，也不得为此提供工具；同时它也切断了作品的可追溯性，在答辩、参展与商用交付时都会构成风险。

拓展阅读

- Sora 技术报告《Video generation models as world simulators》（2024）——读它请带着"哪些是证据、哪些是愿景"的问题意识，并与本章 3.3 节的反证对照。
- 综述《Spatiotemporal Consistency in Video Generation: A Survey》（arXiv:2502.17863）——时间一致性的"官方地图"，本章 3.3 节的两类崩坏就是它的分类框架。
- VideoPhy（ICLR 2025）与 Physics-IQ（Google DeepMind）两篇物理评测论文——"看起来真 ≠ 懂物理"的原始出处。
- 万相 3.0 官方提示词指南与可灵用户指南——把"读官方文档"当作练习：找出它们的提示词结构、参考图上限与时长参数，与本表 3-10 核对。
- 《人工智能生成合成内容标识办法》全文（2025 年 9 月 1 日施行）与 GB 45438-2025——重点读"标识要求"与"禁止行为"两部分，对照本章表 3-18 逐条打勾。
- Variety 对 2026 年圣丹斯 AI 影片的报道——记住其中那位导演的话："You Can't Make a Great AI Film on Your Lunch Hour"（你不可能用午休时间做出一部伟大的 AI 电影）。

第 4 章

三维模型生成与编辑

从一段话、一张图，到一个能拿在手里转的立体物件——中间隔着的不只是"生成"，还有一整套质检与交接。

学习目标

1. **知识：**理解网格、点云、体素、隐式场、神经辐射场、三维高斯泼溅六种三维表征的取舍；讲清 SDS 蒸馏与原生三维生成两条技术路线，以及"重建"与"生成"的区别；理解拓扑、面数、UV、法线、PBR 与导出格式这些工程概念各管什么。
2. **能力：**能用网页工具完成文生三维与图生三维；能对生成结果做多视角质检并填写缺陷记录表；能按"导出检查清单"把资产干净地交给下一个环节（建模软件或游戏引擎）。
3. **素质：**理解"背面幻觉"等失败模式背后的数据根因，养成"不迷信预览图"的判断习惯；认识文化遗产数字化复原中"真实性"的边界与资产许可的法律风险，形成敬畏历史、尊重原作者权益的创作态度。

对应第 5 周 · 3 学时　讲授 2 学时 · 实践 1 学时　主线推进：楼台 / 扁舟 / 玉盘三维初稿

> 小舟在网页工具里敲下一行字：**"唐代临江楼台，木结构，两层歇山顶，水墨风格"**。四十秒后，一个楼台出现在预览窗口里。他转动视角——正面看确实像那么回事：飞檐翘角、木格窗、栏杆都在。可当他转到背面，那面本该是墙的位置糊成了一片起伏的疙瘩，屋檐从某个角度看甚至和柱子"长"在了一起。他把模型导出，拖进引擎，又发现第三个问题：这个楼台比旁边的桥大了近百倍，而且是横躺着躺在地上的。

这不是工具的"故障"，而是这条路线上几乎一定会发生的事。三维生成的门槛，从来不在"能不能生成"，而在"生成之后你怎么看、怎么修、怎么交出去"。这一章要回答三个问题：**一、**三维数据在计算机里到底是怎么表示的，为什么"没有一种表示是万能的"？**二、**从一张图或一段话到一个三维模型，AI 走的是哪条路，为什么会"背面幻觉"、会面数膨胀、会把薄壳物体算成一团？**三、**拿到模型之后，拓扑、面数、UV、PBR、导出格式这些名词各自管什么，一份"导出检查清单"能替你挡掉多少坑？学完这一章，你会得到一种新的眼光：**看一个三维资产，不看它的正面有多漂亮，而看它的背面、它的线框、它的数据。**

📌 本章路线图

4.1 三维表征家族（网格 / 点云 / 体素 / 隐式场 / NeRF / 3DGS）→ **4.2** 两条来路：重建与生成 → **4.3** 让 AI"想"出三维：SDS 蒸馏与原生三维生成 → **4.4** 纹理与多视角一致性 → **4.5** 拓扑 · 面数 · UV · PBR → **4.6** 导出、交接与检查清单 → **4.7** 工具地图 → **4.8** 失败诊所 → **4.9** 动手实践 → **4.10** 关键概念表 → **4.11** 课程思政 → **4.12** 本章产出与下一站。

与前一章的关系：第 2 章你做的是"平面上的画面"，第 3 章让它动了起来；这一章要把画面变成**有体积、可旋转、能被摆进场景**的东西——第 7 章搭夜江场景时的每一件道具，都要从这一章的产出里来。

4.1 三维数据在计算机里长什么样：表征家族

先解决一个最基础、也最容易被跳过的困惑：计算机里根本没有"立体"这个东西。屏幕是平的、显存是线性的、文件是字节的。所谓"三维模型"，只是**用一种数据结构去描述一件东西的形状**，再由渲染器把这份描述"翻译"成平面图像给你看。用什么数据结构去描述形状，就是本章的第一个核心概念：**三维表征**（3D representation）。

你可能会想：那还不简单，把形状直接记下来不就完了？问题在于"形状"这个东西太自由了——它可以是一个光滑的球，也可以是一片薄到几乎不存在的树叶，还可以是一团烟雾、一件镂空的铜器、一栋内部有房间的建筑。没有哪一种记录方式能同时把这些都记好。业界主流的六种表征，各自擅长一部分、吃亏一部分。记住这个结论：不存在"最好的三维表征"，只存在"对某个用途最划算的表征"——这个认知，比记住六个名词重要得多。

图 4-1　三维表征家族对照。六种"记录形状"的方式，各自像一个生活物件。这张图是本章的地图：后面讲的每一种技术，都建立在其中某一种或某几种表征之上。

下面把六种表征逐个讲清楚。每一段都按同一个套路：**它是什么 → 像什么 → 为什么有人用它 → 它吃亏在哪 → 在你的短片项目里什么时候会遇到它。**

◆ 点云：把形状记成"一堆点"

点云（point cloud）是最直觉的一种：在物体表面撒一把"没有体积的点"，每个点记录自己的三维坐标，还可以顺带带颜色、法线方向等信息。激光雷达、深度相机、摄影测量软件输出的第一份成果往往就是点云。

它的好处是"采集即所得"——扫一圈，点云就有了，不需要任何拓扑结构。它的麻烦也来自同一处：点与点之间没有任何连接关系。计算机不知道哪些点是"同一个面的邻居"，所以渲染时你看到的是无数小点；点不够密，物体上就会出现空洞；想把它变成能放进引擎的东西，通常还得走一步"把点连成面"。另一个工程麻烦是稀疏且不规则——想高效渲染与压缩，往往要做复杂的预处理。这也是第 7 章拼场景时会遇到的现实问题。

在课程里，点云最可能出现的位置是"扫描复刻"环节：小舟如果想用手机把一座真实的石桥扫下来，得到的第一份数据就是点云。

◆ 网格：把形状记成"一张张三角形的小板"

网格（mesh）是当下最主流、也是你将来打交道最多的一种表征。它的思路是：用许多小平面把物体表面"蒙"起来。每个面通常由三个顶点组成一个三角形（也有四边形），顶点被相邻的面共享。于是"形状"变成了三份清单：**顶点表**（有哪些点、坐标各是多少）、**面表**（哪几个点连成一个面）、**顶点属性**（法线方向、UV 坐标等）。

为什么它是主流？因为工程生态最完整：内存高效、GPU 光栅化渲染极快、几乎所有建模、动画、雕刻、游戏引擎工具链都为它而建，格式标准也最成熟（`.obj`、`.fbx`、`.glb`）。

它吃亏的地方同样清楚：**第一**，它要求"干净的几何"——如果网格里有空洞、有内外穿插的面、有反了的法线，很多下游工序会直接出错；**第二**，精度是靠"堆面数"换来的，越精细越费内存与算力；**第三**，模糊或透明的区域（火焰、薄纱、玻璃）用面片很难表达好；**第四**，它天然不带"视角相关效果"（比如丝绸随角度变化的光泽），需要额外的贴图技术来补。

一句话总结它在本书里的地位：AI 生成的最终交付形态，通常就是网格——你在网页工具里点"导出"，拿到的基本都是网格文件。因此本章后半部分几乎都在讲"怎么把一个 AI 生成的网格，修成能用的网格"。

◆ 体素：把形状记成"一堆同样大小的方块"

体素（voxel）是三维版的像素：把空间切成一张规则的立体方格网，每个小格记录"这里是实体还是空的"（也可以记录颜色、材质等）。你可以把它想成"用同样大小的方块填满一块豆腐"——内部与外部的区别天然被表达出来了。医学影像（CT、MRI）的处理、体积雾与参与介质的建模，都是它的传统地盘。

它的致命伤可以用一道小学算术说清：边长减半，方块总数变成 8 倍。分辨率稍微提一档，数据量就爆炸；而空间里大部分格子其实是空的（物体只占一小部分表面附近），这些空格子是纯粹的冗余。此外，方块只能表达方块，表面看起来会有明显的"块状伪影"，想要光滑还得靠插值或后续的表面提取。

体素在本章的重要性不在"直接用"，而在"间接用"：它是很多现代方法（稀疏体素、混合表征）的内部工作格式，也是你理解"分辨率为什么会吃内存"的第一课。

∫ 深入一格 · 为什么"细一格"要付出八倍代价 选学 · 数学直觉

把一块 1 米见方的立方体切成小方格，如果每边切 **n** 段，总格数就是 n×n×n。这是三维与二维最根本的不同：二维是平方增长（每条边翻倍，总面积变 4 倍），三维是立方增长（每条边翻倍，总体量变 8 倍）。

$$格数 = n \times n \times n = n^3$$

边长减半意味着 n 翻倍（n 从 2 到 4），总格数变为 $4^3 \div 2^3 = 8$ 倍；若每边细分到 4 倍，则总量变为 64 倍。

这个 n^3 不只是"内存问题"，它决定了一整条技术路线的命运：为什么体素方法长期只能做小分辨率？为什么要发明"稀疏体素"（只记录表面附近的格子）与"八叉树"（空的地方用大格子概括）？为什么"隐式场"看起来优雅——它用一组参数而不是一堆格子描述连续表面，天然绕开了 n^3？你在 4.3 节会看到，2026 年一批领先的原生三维生成模型，恰恰是在"稀疏 + 层次化"这条路线上取得突破的。

◆ 隐式场与 SDF：把形状记成"一张离表面多远的隐形标尺"

隐式场（implicit field）的思路完全反过来：不记录"表面在哪里"，而是记录"空间里每一个位置，距离表面有多远、是在里面还是在外面"。最常用的具体形式是**有符号距离场**（signed distance field，SDF）：给空间每个位置一个数——在物体外部是"到表面的距离"（正数），在内部是"到表面的负距离"，恰好落在表面上就是 0。于是表面被定义成"所有函数值等于 0 的位置的集合"，这条界线在数学上叫等值面。

它的迷人之处在于"光滑"与"可运算"：因为描述表面的是一个连续函数，它天然光滑、没有锯齿；而且两个形状可以做**布尔运算**（取并集、交集、差集）——"在楼台上开一个门洞"在隐式场里就是两个场做一次数学运算。它数值上也稳定，适合多视图融合与实时建图；NeuS、VolSDF 这类方法能用隐式场做高精度表面重建，最后用 marching cubes（行进立方体，一种把隐式场转成三角网格的经典算法）把等值面"提取"成网格。

代价有两条：**一是"看不见"**——隐式场本身没有可视化形态，要看到它必须先把空间离散化、取样，再提取表面，这一步在高分辨率下存储与算力开销极高；**二是渲染方式慢**——体积渲染要沿光线一步步前进（后文 NeRF 会细讲），在边缘设备上很难跑到交互帧率。所以它更适合"高精度重建与分析"，而不是"游戏里实时跑的场景"。

课程情境：小舟要做一个"从几十张照片重建的楼台"，重建算法内部很可能就是隐式场；但他最终拿去用的，一定是它提取出来的网格。

◆ 神经辐射场 NeRF：换一个问题问"这里是什么"

神经辐射场（neural radiance field，NeRF）是 2020 年之后最有影响的一类三维表示。它的想法很别致：不去描述"表面在哪里"，而是回答"沿着某条视线看过去，你看到什么颜色"。

具体做法：训练一个小型神经网络，输入是"空间中的一个位置 + 一个观察方向"，输出是"这里的体密度（有多不透明）与颜色"。渲染一张图时，把相机出发的每条光线当成一条射线，在射线上密密麻麻采样若干点，每个点都去问一次神经网络，再把沿途采到的颜色与密度按权重累积起来，就得到这个像素的颜色。**训练方式**是：从多张已知相机位置的照片出发，让渲染结果与真实照片一致，反向修正网络参数——直到这个"可查询的场"能重现所有视角的观感为止。

它的优势很迷人：视觉保真度高，还能表达"视角相关外观"（同一处表面，换个角度看颜色会变——这正是金属、丝绸、漆器这类材质的真实行为），连毛发、软阴影这类细节都能捕捉。纯 MLP 版本模型也很紧凑。而 2022 年之后的 Instant-NGP 用哈希网格加速，把训练从数小时压到秒或分钟级，并支持实时渲染。

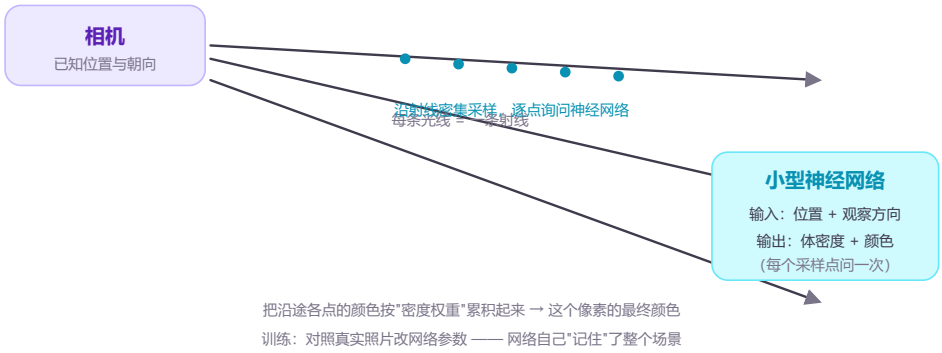


图 4-2　一根光线穿过场景。NeRF 的渲染是"沿视线采样 + 加权累积"，训练是"渲染结果对照片"。理解了这两句话，你就理解了它为什么真实，也为什么慢、为什么难编辑。

NeRF 的短板几乎与它的优点对称：**训练与推理昂贵**（原生版本训练要数小时）；**难编辑**——因为它把所有东西揉进了一个神经网络，想"把楼台移走"往往只能重新训练；**渲染慢**；而且它没有显式的表面几何，游戏引擎拿不到"可碰撞的墙面"，物理系统不知道哪里能走、哪里会撞。业界对它的普遍定位是：离线或分阶段的工具，不适合实时 SLAM、机载避障或亚秒级更新的场景。混合结构（显式体素、哈希、点结构）能大幅加速，但在大场景下又容易把内存撑爆。

◆ 三维高斯泼溅 3DGS：一团团会渲染的"彩色棉花糖"

三维高斯泼溅（3D Gaussian Splatting，3DGS）是 2023 年之后迅速走红的一类表示。它从一个非常务实的问题出发：既然要从照片重建场景，渲染要快、训练也要快，那能不能不用神经网络、直接摆一堆"半透明的彩色小球"？

它用成千上万个"三维高斯"来代表场景——你可以先粗略地理解成一个个位置、朝向、大小各不相同、还半透明的椭球云团。每个云团有自己的颜色、透明度、空间姿态；渲染时把它们"泼溅"（splatting）到屏幕上叠加成画面。它的优点非常诱人：**桌面 GPU 上能做到 100 帧以上实时渲染**，训练速度快（与 Instant-NGP 相当），而且因为是显式的（一堆有参数的对象），可以做剪枝、变换、合并等"可编辑"的操作。

它的短板同样鲜明：会产生**悬浮物**（floaters）——飘在空中的半透明团块；表面质量也常被诟病。一项 2026 年的跨表征标准化对比给出了一句相当刺耳的结论：**非结构化的 3DGS 视觉分数最高，底层几何却最差**；而带显式空间结构的技术（体素、混合体素-高斯）在拓扑连贯性与抑制悬浮伪影上明显更好。此外，高斯图元数量常达十万到千万级，存储随数量增长。

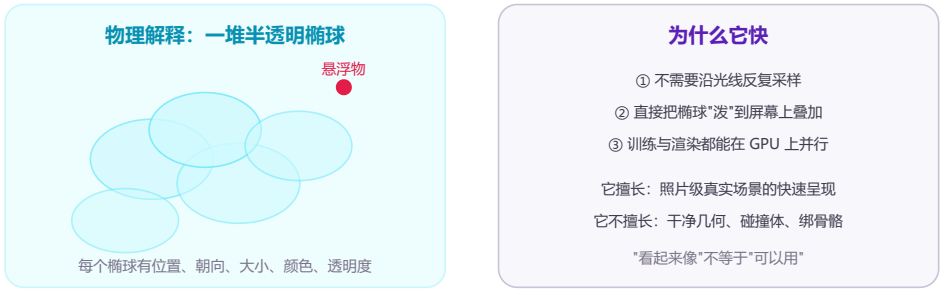


图 4-3　三维高斯泼溅。把场景交给一团团半透明椭球，换来极快的渲染；代价是它的几何不能直接当墙用。第 7 章搭场景时，"3DGS 背景 + 网格道具"会是常见的混合方案。

还有一个正在发生的变化值得知道：**3DGS 正在被"标准化"。**据检索到的信息，OpenUSD 已批准高斯泼溅支持，Khronos 也提出了 glTF 的候选扩展（`KHR_gaussian_splatting`），互操作性正在改善（覆盖 Vision Pro、Omniverse 与主流引擎）。对创作者而言，这意味着 3DGS 不再只是"实验室玩具"，而可能成为资产交付的一等公民。同时也要知道它的另一面：GS 的专业服务成本目前约为标准摄影测量费率的 1.5 倍。

表 4-1　六种三维表征的能力权衡（"没有最好，只有取舍"）

表征	一句话定位	主要优势	主要劣势	典型用途
点云 point cloud	一堆带坐标的点	结构简单、采集端直接产出、每点可带颜色与法向	无连接关系与拓扑；渲染有空洞；稀疏不规则，压缩与高效渲染要复杂预处理	扫描第一步、测量、体积计算
网格 mesh	三角形拼成的外壳	内存高效、渲染快、工具链最成熟、格式标准、能绑骨骼	需要干净几何；精度靠堆面数；动态拓扑一致性难维护；模糊与透明区域表现差	游戏、动画、影视、3D 打印
体素 voxel	同样大小的立体方块	天然表达内部与体积结构，处理简单	存储随分辨率三次方增长；稀疏区域冗余巨大；块状伪影	医学影像、体素美术、内部结构建模
隐式场 / SDF implicit field	一张"距表面多远"的连续函数	表面光滑连续、拓扑灵活、可做布尔运算、数值稳定、支持精确表面重建	存储与可视化必须离散化取样；高分辨率成本极高；体积渲染难达交互帧率	高精度重建、逆向工程、布尔造型
神经辐射场 NeRF	沿视线查询的"可渲染函数"	视觉保真度高、含视角相关外观、能捕捉细节、从多视角照片学习	训练与推理昂贵、难编辑（须重训）、体积光线步进慢、无显式碰撞几何	离线高真实感重建、虚拟导览、展示
三维高斯泼溅 3DGS	一团团半透明椭球	实时渲染（桌面 GPU 100+ FPS）、训练快、显式可编辑（剪枝 / 变换 / 合并）	悬浮物与表面质量问题；存储随高斯数量增长（常 10 万–1000 万）；几何不可直接使用	照片级真实场景、网页嵌入、虚拟导览

把这张表再往上一层看，一项 2026 年的跨表征标准化对比（覆盖摄影测量、纯 SDF、非结构化 3DGS、混合体素-高斯、显式稀疏体素五套框架）给出了几条很有价值的结论，值得你记住：

- **传统摄影测量的几何精度最高**（因为它强制多视角的刚性几何约束），但在**没有纹理的区域会失败**——白墙、纯色桌面这类"看起来没东西"的地方最容易出错。
- **渲染类方法里，非结构化 3DGS 视觉分数最高、几何最差**；带显式结构的技术（体素、混合体素-高斯）在拓扑连贯性和抑制悬浮伪影上明显更好。
- 结构化体素还能**避开"从混乱无组织的图元里提取几何"带来的严重内存瓶颈。**研究者的结论是：把显式空间结构整合进可微渲染管线，对"可扩展 + 几何准确"的大规模重建非常重要。
- 再加一条"时间维度"：**4DGS** 在 3DGS 上增加时间维，时序一致性强、压缩高效、支持流式重建，但形变处理复杂、内存开销高。

还有一组"管线与集成"的差异，解释了三者在实际生产中的区别——它回答的是"我要用它，得先做多少准备、之后能改多少"：**网格**的预处理要经过摄影测量式的完整链条（相机标定 → 稠密点云 → 网格化），建模人工量最大，但可编辑性最好、引擎与 XR 集成完全支持，只是照片真实感是"中到高"（它没有视角相关效果）。**3DGS** 的预处理轻得多（只需稀疏点云），几乎不需要人工，照片真实感高到极高，可编辑性中等（在"泼溅"这一层级上改），进引擎要靠自定义着色器。**NeRF** 预处理最轻，几乎不用人工，真实感也高，但可编辑性最低（基本等于重训），而且引擎不直接支持——它没有显式表面，物理与碰撞无从谈起。这句话你可以当作选型的第一原则：**要"能撞、能走、能被摆"的场景，用网格；要"看起来真"的背景，才考虑 NeRF 或 3DGS。**

> **🎯 表征选型的三个问题**
>
> 面对一个任务，先问三句话再选表征：① 它需要"能交互"吗（碰撞、绑骨骼、被物理系统处理）？需要 → 网格。② 它更需要"看起来真"还是"量得准"？看起来真 → NeRF / 3DGS；量得准 → 摄影测量或 SDF 重建。③ 后面还要改吗？要改 → 显式表示（网格、点云、3DGS）；不用改只看 → 隐式与 NeRF 也可以接受。本书的贯穿项目是"能摆进场景、能出动画"的短片，因此每一件资产的最终形态都必须是网格——这是我们后面所有讨论的前提。

最后补一句"前瞻"：三维表征的研究并没有停在六选一。今天的前沿在做的，恰恰是**把不同表征的优点拼起来**——比如"3DGS + 网格"的混合表征（既能实时渲染照片级场景，又能导出可进引擎的网格），或者"稀疏体素 + 隐式解码"的组合。你会在 4.3 节看到，2026 年领先的原生三维生成模型，很多就建立在稀疏体素这一类显式结构上。表征的融合，是这一轮技术进步的一条暗线。

4.2 两条来路：重建与生成

知道了三维"长什么样"，下一个问题是：它从哪来？追根究底只有两条来路——**重建**（reconstruction）与**生成**（generation）。这两条路经常被混为一谈，因为它们都用 GPU、都产出网格文件、都能在网页上转着看；但它们的**输入、目标、评价标准、失败方式**完全不同。把这条界线分清，你才能在选工具时不迷路。

◆ 4.2.1 重建：把真实世界里已经存在的东西"搬"进来

重建的输入是对一个真实物体的观测——一圈照片、一段环绕视频、几次激光扫描，甚至医院的 CT 序列。它的目标是：**在计算机里复现这个真实存在的东西**。它有"标准答案"——那个真实物体本身，所以它可以被客观地度量：几何误差多少毫米、渲染结果与照片差多少分贝。

传统重建的主力是**摄影测量**（photogrammetry，也叫运动恢复结构 + 多视图立体视觉，MVS）工艺：拍一圈照片 → 相机标定（算出每张照片拍摄时的位置与朝向）→ 稠密点云（算出表面每个点的三维坐标）→ 网格化（把点云连成面）。这套链条成熟、稳定、精度高——4.1 节提到的跨表征对比里，传统摄影测量的几何精度是全场最高的，因为它强制多视角的刚性几何约束，每个点都被多张照片"交叉验证"过。

它吃亏的地方同样明确：**没有纹理的区域会失败**。白墙、纯色桌面、抛光金属这类表面，算法找不到足够的特征点去"对上"，重建立刻出现空洞或起伏——这一条你务必记住，它是后面"缺陷诊所"里好几类问题的同源病因。此外，透明与反光物体（玻璃、镜面、水面）对摄影测量也是长期难题，因为它们的光学表现不符合"表面同一个点在任何视角下颜色相同"的假设。

AI 时代给重建加了两台"新引擎"：**神经隐式场重建**（NeuS、VolSDF 一类，把隐式场当成可优化的函数，训练完再提取表面）和**三维高斯泼溅重建**（3DGS，从照片直接摆出一团高斯）。它们把重建的"可处理场景"大幅拓宽——稀疏视角、弱纹理、遮挡区域都能恢复，视觉真实感也高。代价是：它们需要显式结构约束才能拿到干净、可编辑、可进引擎的几何；单纯的高斯云"看起来像"但"不给你墙面、不给碰撞"。还有一条前沿：**4DGS** 在 3DGS 上加了时间维，可以重建"会动的东西"（挥舞的手臂、流动的水面），时序一致性强、支持流式重建，但形变处理与内存开销都更重。

◆ 4.2.2 生成：造一个世界上原本没有的东西

生成的输入是"描述"——一段文字、一张参考图、一幅手绘；它的目标是：**合成一个此前不存在的三维物体**。注意这句话的分量：世界上没有一个"标准答案"能拿来对照。所以生成的评价标准天然是**主观**的——"看起来合理吗？""像不像提示词说的东西？""比例对不对？"这些判断要么靠人打分（人偏好类评测），要么靠模型打分（自动化指标），而 4.4 节你会看到，这两把尺子量的东西并不总是一回事。

生成的历史，其实是一部"信息不足时如何补脑"的历史。设想你只给 AI 一张楼的正面照片，要它产出完整的三维模型。正面照里根本没有背面的任何像素——AI 必须"猜"。那么它根据什么猜？根据它在海量数据里见过的所有楼。这与人类的行为一模一样：你从没去过那座楼，但你能凭经验想象"背面大概还有一面墙、还有窗"。这个类比就是理解生成的核心：生成模型的一切长处与短处，都来自"它见过什么"和"它没见够什么"。

> **💡 一句话分界线**
>
> 拿到任务先问一句："**世界上有这个东西吗？**" 有，而且我能拍到它 → 走**重建**，你的敌人是"观测不足与噪声"。没有，或者拍不到 → 走**生成**，你的敌人是"模型没见过"。**这条分界线，决定了你后面所有的质检动作该怎么做：重建的质检是"对表"（与真实物体比），生成的质检是"找茬"（凭常识找不合理处）。**

图 4-4　两条来路。左路有标准答案，可以量化；右路没有标准答案，只能由人（或模型）来评。本书的贯穿项目两种都会用到——檐角的石狮雕像可以重建，而"江心最特别的那座楼台"只能生成。

表 4-2　重建与生成：五处关键差别

维度	重建 reconstruction	生成 generation
输入	真实物体的多视角观测（照片 / 视频 / 扫描）	描述（文字 / 参考图 / 手绘）
目标	复现已存在的东西，越忠实越好	合成不存在的东西，越合理越好
有没有标准答案	有：真实物体本身	没有：只有"像不像 / 好不好"的主观判断
评价方式	几何误差、与照片的渲染一致性等客观指标	人偏好投票、自动化质量指标（两者会打架）
典型失败	弱纹理处空洞、透明反光物体崩坏、遮挡区缺失	背面幻觉、多面神、面数膨胀、薄壳崩坏

◆ 4.2.3 中间的灰带：信息量决定你站在哪条路上

现实中两者的分界线并不像表格那么清爽，因为真正的裁判是"信息量"。给 AI 一百张环绕照片，它几乎不需要"想象"，重建就能解决；给 AI 一张正面照片，它必须大量"想象"，问题立刻变成生成问题。

这里藏着本章第一个原理性的关键点——**单视图的欠定性**。研究者对"只给一张正面图，模型为什么必然搞错背面"做过形式化分析：单视图条件本身就不足以确定背面的几何，这不是某个模型的 bug，而是问题的数学性质。有研究用视觉语言模型往生成流程里注入"常识"来改善背面结构，把背面几何的 Chamfer 一致性最多提高了约 37.2%——注意这个数字的含义：注入知识有效，但达不到完美，因为知识再多也变不出观测。你会在 4.8 节看到这类误差在真实工具里的样子：从单张正面图生成的物件，背面常常"糊成一团"或"长出不该有的结构"，社区里把这类产物叫作模型的"隐藏伪影"。

于是"中间地带"的工程策略就清晰了：**能多给一个视角，就多给一个视角。**同一个工具，给它一张正面图与给它"正面 + 侧面 + 背面三张图"，输出质量的差距常常大过一个工具的换代差距。这也是本章实践环节的第一条纪律：**生成前先问"我手上有几个视角"，而不是先挑工具。**

贯穿项目里的三个道具，恰好落在谱系的不同位置，小舟他们的选择也就自然分开了：**楼台**——"唐代临江楼台"世界上并不存在确切的对应物，走**文字生成**；**扁舟**——第 2 章已经画出了水墨风格的分镜，小月又从参考里挑了一张侧视角度最清楚的画面，走**图像生成**；而檐角那一尊小石狮，如果用手机围着真实雕塑拍一圈，就是一次标准的**重建**。至于**玉盘**——一个光滑的圆盘——根本不必生成，见 4.9 节的实践二，那里藏着一个"不要什么都交给 AI"的反向教学点。

4.3 让 AI"想"出三维：从 SDS 蒸馏到原生三维

4.2 节讲了"生成"是什么，这一节要回答"生成是怎么做到的"。这条路线上有一段非常精彩的技术史：**短短四年，业界换了两次底层范式。**理解了这两次换档，你对任何"AI 生成三维"工具的判断力都会高一个层次——因为它们的短板、速度、可用性，全部继承自它们所走的那条路线。

图 4-5　三维生成的四个世代。从"显式离散"到"神经隐式"，再到"借 2D 模型指导 3D"，最后到"直接生成本原三维"。每一代都解决了上一代的一个死结，也留下了自己的软肋。

◆ 4.3.1 第一次换档：用 2D 模型当"老师"——SDS 蒸馏

2022 年前后，二维图像生成模型（第 2 章的主题）已经能"听懂人话画画"了。一个顺理成章的念头出现了：**二维模型懂那么多，能不能把它的知识"蒸馏"给三维？**

分数蒸馏采样（score distillation sampling，SDS）就是这套思路的代表。它的流程像一场"雕塑家与评委"的循环，理解它只需要看懂四步：

1. **先捏一个坯子。**随机初始化一个三维表示——原始版本用的是 NeRF（一个随机初始化的神经辐射场）。
2. **从随机视角"拍一张照片"。**把这个三维表示渲染成一张二维图。
3. **请二维扩散模型当评委。**把渲染图与提示词一起交给预训练的二维扩散模型，让它判断："这像不像'一只柯基'？该往哪个方向改才更像？"扩散模型给出的是一个"梯度的方向"，也就是"这样改会更像"的信号。
4. **按评委意见改坯子。**把这份"二维的修改意见"翻译成对三维表示的参数更新——分数蒸馏采样——然后回到第 2 步，换一个新视角再来一遍。循环几百上千次，坯子逐渐长成评委认可的模样。

SDS 有一个足以写进教科书的名字来源：DreamFusion，2022 年提出。它的历史地位在于"第一次证明了：不需要任何三维训练数据，仅靠二维生成模型，就能造出三维物体"。这在当时非常震撼。但它也让所有人看清了这条路的天花板。

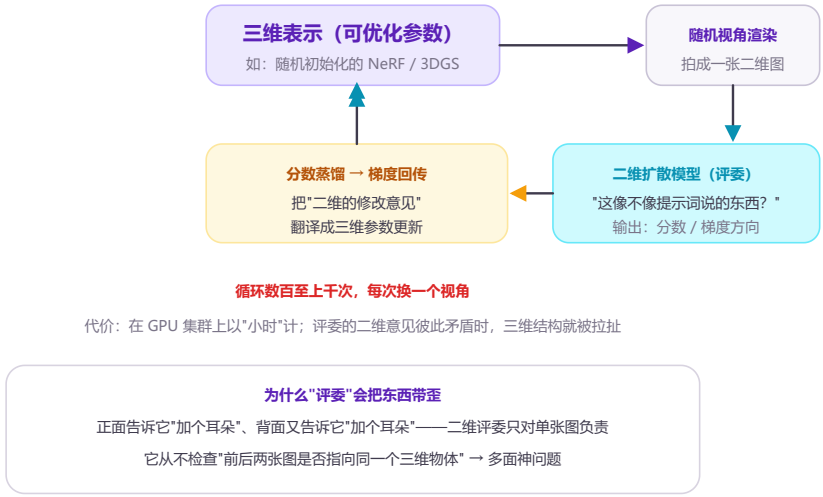


图 4-6　SDS 的"雕塑家与评委"循环。真空中看着天经地义，现实中评委只对单张二维图负责——这正是"多面神"与背面幻觉的温床。

第一类问题：多面神。因为你让二维评委每张图分开打分，它从不检查"正面的狗和背面的狗是不是同一只"。于是模型学到的东西很可能变成：每个方向看过去都像一只完整的狗脸——转一圈，长着三张脸。这个现象用罗马神话里双面的门神雅努斯命名，叫**多面神问题**（Janus problem）。它是背面幻觉的"近亲"：两者的根源相同——二维监督信号不包含"三维自洽"这一要求。

第二类问题：慢。最初的 DreamFusion 生成一个三维模型要**约 6 个 TPU 小时**——不是六分钟，是六个小时。同期把它加速的 Magic3D 降到了**约 5.3 个 A100 小时**，依然是"睡一觉再来看"的量级。对研究者，这是可以忍受的；对"给短片赶工的小组"，这是不可接受的。整个 SDS 生态后面大量的精力都花在提速上：TP-SDS、DreamFlow、PFD 这类改进让蒸馏速度提升了约 3.3–4.8 倍（如 DreamPropeller 约 3.28 倍），也有一批工作（Gsgen、Magic123）把坯子从 NeRF 换成 3DGS 来加速渲染侧。速度确实改善，但"小时级"与"秒级"之间的鸿沟并没有被填平。

第三类问题：监督失配与分布外输入。评委是在"互联网二维图片"上训练的。它觉得"好看"的东西，不一定在三维上成立：一根细杆在它的认知里可以悬空、一片薄布可以无厚度。而当你要生成的东西超出它的经验（现实中少见的机械结构、特殊的榫卯、某种只在本地见过的器物），它给出的"修改意见"就更容易跑偏。用报告里的话说：SDS 的问题可以归纳为**收敛缓慢、监督信号失配、与训练分布外的输入**三类——三类都源自同一个事实：**二维老师教不出三维的严格自洽**。

表 4-3　SDS 路线的三个代表：一代比一代快，但都在"小时级"

代表工作	核心改进	生成耗时（代表性报告）	质量参考（3D-FID）	多面神比例（评测口径）
DreamFusion（2022）	首次提出 SDS：用 2D 扩散模型蒸馏出三维	约 6 TPU 小时	—	约 35.71%
Magic3D	粗到细的两阶段：先粗网格再精修	约 5.3 A100 小时	75.96	—
ProlificDreamer	改用变分分数蒸馏（VSD）缓解过饱和与过平滑	仍为小时级	58.75（越低越好）	约 40.00%

把这些数字放在一起看，一个结论不言自明：**SDS 的改进是真实的，但它作为"生产工具"的气质没有变**——慢、易幻觉、几何质量不稳定。它是三维生成史上绕不过去的里程碑，却不是小舟今晚要用的工具。**注意表里那个耐人寻味的事实：换用 VSD 后质量指标明显改善，多面神比例却没有自动消失**——这再次证明"二维好看"和"三维自洽"是两把尺子。本章后面的失败诊所会反复回到这个主题。

◆ 4.3.2 第二次换档：原生三维生成——直接在三维里"砌"

2024 年之后，路线换了。新思路不再"用二维老师教三维学生"，而是**直接训练懂三维的大模型**：把三维数据切成"积木"（结构化潜变量），让模型学会"哪些积木该拼在一起"，然后像第 2 章讲画图那样，从噪声出发一步步"去噪"出这些积木，最后解码成网格。这条路叫**原生三维生成**。

它的关键突破点有三个，值得逐个弄明白——它们解释了为什么这条路能"秒级出模"：

- **把三维切成"够小又够用"的积木。**在 1024^3 这样高的分辨率上，如果把每个格子的状态都当成 token 交给 Transformer，序列长度会爆炸（有稀疏表示在同样分辨率下要处理约 **22.5 万**个 token）。解法是学一个"压缩器"：例如 TRELLIS.2 的 **SC-VAE** 把 1024^3 的分辨率做 **16 倍下采样**，压缩到**约 9,600 个 token**——降了两个数量级。序列短了，注意力机制才跑得动。
- **用能扛任意拓扑的表示。**老式体素只能表达"方块"，遇到开放曲面、非流形结构、带内部空腔的物体就露怯。新方法用稀疏的、贴着表面的结构化体素（如 TRELLIS.2 的 **O-Voxel**），只记录表面附近的信息，能表达开放面、非流形与封闭体等任意拓扑。
- **用生成式大模型"去噪"出积木。**主体是一个多段式的流匹配 Transformer：TRELLIS.2 的做法是三个各约 **13 亿参数** 的阶段，由粗到细逐步生成。流匹配（flow matching）与扩散模型是同一族思路——从噪声出发、一步步"搬"到目标分布。

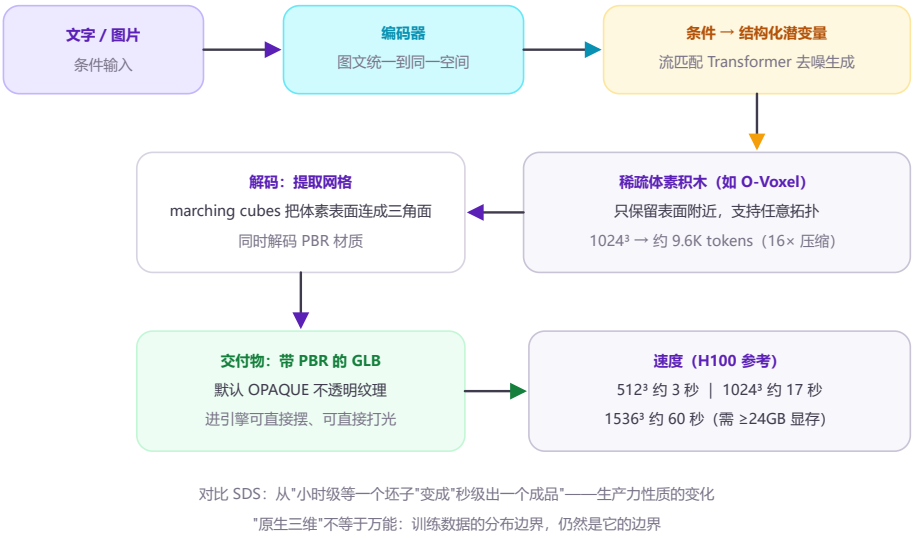


图 4-7　原生三维生成的流水线。压缩（SC-VAE）→ 稀疏体素表示（O-Voxel）→ 流匹配去噪 → 解码出网格与材质。三段式设计让"秒级生成"第一次成为现实。

以 TRELLIS.2 为代表的新一代模型有多能打？把它公开的关键数据摆出来：约 **40 亿参数**；输出是**带 PBR 材质的 GLB**（可直接进引擎打光）；采用 **MIT 许可**（开源协议宽松，商用友好——对照 4.11 节你会看到，这在同行里并不常见）；全表面检验的 F1 达到 **0.971**（上一代 TRELLIS 仅 0.074，这是一个数量级以上的跨越）；法线方向的 PSNR 达到 **43.11 / 35.26 分贝**（越高越好）；用户偏好测试中拿到 **66.5% / 69.0%** 的胜率。代价是需要 **NVIDIA 显卡且显存不低于 24GB**——这条硬件门槛意味着小组里"笔记本电脑本地跑"基本出局，要么用在线工具，要么租云 GPU。

不过要给你一句研究视角的提醒（这条判断来自本地研究报告，请把它和厂商的宣传语区分开）：TRELLIS.2 是 2026 年被"最广泛复用"的三维生成骨干，但"公认最强"这类说法是被夸大的——同期的混元 3D 系列是完全平行的另一条骨干家族，两者各有拥趸，选择取决于场景与许可条件，而不是一个排行榜。研究者还指出：2024–2026 的另一个方向是**融合**——一批新工作（Mix3R、RecGen3D、Pixal3D 等）反过来从重建领域"借"机制（多视角约束、几何先验），说明"生成"与"重建"两条河正在汇合。这印证了本章 4.1 的立场：**表征与方法都在向"取长补短"演化，不要押注单一教条。**

表 4-4　2026 年原生三维生成的代表模型（数据取自各来源公开材料，标注口径见正文）

模型	路线与规模	关键能力	使用门槛 / 备注
TRELLIS.2	结构化潜变量 + 稀疏体素（O-Voxel），约 4B 参数	任意拓扑、PBR GLB 输出、H100 上 1024^3 约 17 秒、MIT 许可	需 NVIDIA ≥ 24GB 显存；适合自建管线的小组
混元 3D 3.1	腾讯出品，云服务为主	每日 20 次免费生成、默认导出 GLB、面数可配置（约 5 万–150 万）、水密参数可选	云 API 有免费积分额度；注意许可条款（见 4.11）
混元 3D 2.1（开源版）	两段式：Shape 3.3B + Paint 2B	本地可跑：形状约需 10GB 显存、贴图约需 21GB 显存	腾讯社区许可（非 MIT）；下载量巨大，社区生态活跃
Tripo Algorithm 3.1	超 2000 亿参数（厂商口径）	导出格式最全（OBJ/FBX/GLB/GLTF/USD/USDZ/STL/3MF）、最高 200 万多边形、8K 贴图、Turbo 约 10 秒	免费额度各来源说法不一（200 或 300 积分/月，见 4.7）；免费层输出 CC BY 4.0
Rodin Gen-2 / Gen-2.5	10B 参数级（Gen-2）	面向游戏与工业；4K PBR；千万级多边形上限	免费生成、付费下载；强制 T/A-pose；STL 通常需修复
Meshy 6	面向 3D 打印与游戏的两条产线	原生四边形重拓扑；水密几何宣称约 97% 通过切片器；默认约 3 万面	免费 100 积分/月（CC BY 4.0 需署名）、每月 10 次下载

把这一节收拢成一段给创作者的话：**SDS 教会了业界"三维可以无中生有"，原生三维教会了业界"无中生有可以又快又好"。**但两次换档都没有解决一件事——模型只会按它见过的分布做事。这意味着你完全可以用今天的工具在几十秒内得到一个"第一眼惊艳"的模型，也意味着你必须学会 4.8 节那一整套"找茬"的本事，否则你会把幻觉当成作品、把 50 万面当成细节、把预览图当成成品。技术越强，质检越贵——这是本章最想刻进你肌肉记忆的一句话。

4.4 纹理与多视角一致性：为什么"正面好看"不可信

到这里，你的模型已经有了"形状"。但对创作者来说，形状只算完成了一半——观众看到的是颜色、质感与光，不是线框。给模型"穿上皮肤"这一步叫**纹理**（texture），它听起来像收尾的美工活，实际上是三维生成里最容易翻车的一段。为什么？因为纹理比几何多背了一个任务：它不仅要"对"，还要在**所有视角下同时对。**

◆ 4.4.1 纹理从哪来：三条路

今天的三维工具给模型"上皮肤"主要有三条技术路线，你认得出它们，就能预判各自的毛病出在哪：

- **联合生成。**原生三维模型在生成几何的同时，一并解码出材质——上一节的 TRELLIS.2 输出的就是"带 PBR 材质的 GLB"。这条路最省事，几何与贴图出自同一次推断，天然自洽；缺点是贴图的分辨率与风格上限受模型控制，不是你想改就能改。
- **先白模，后投影。**先生成无纹理的"白模"，再从多个视角渲染它、用二维生成模型逐视角补画颜色，最后把各视角的颜色**反投影并烘焙**回三维表面。技术上灵活，可以用第 2 章的图像模型资源来"画皮"；风险也集中在这一步：不同视角的图像并不总是互相一致，重投影时会在接缝处出现色差、错位、拖影。
- **图生三维自带纹理。**你给的那张参考图本身就带颜色，工具把它"贴"到推断出的几何上。这条路的问题最直白：参考图只有一面，其余各面的颜色全靠模型脑补——于是"正面是青瓦、背面糊成一片灰"就成了高发缺陷。

表 4-5　纹理生成的三条路线与各自的软肋

路线	大致做法	优势	软肋（质检时重点看）
联合生成	几何与材质在同一次推断中解码（如 PBR GLB 直出）	几何贴图自洽、交钥匙式产出	分辨率与风格受模型约束；纹理细节可能偏"平均脸"
先白模、后投影	多视角渲染 → 逐视角上色 → 反投影烘焙回三维	可借用二维图像模型的强大画风	视角间不一致会在接缝处暴露；拖影、错位、色差
图生三维自带纹理	参考图的颜色投影到推断出的几何上	一键完成，风格忠实于参考图	只看过一面：背面颜色与细节靠猜

◆ 4.4.2 被"烤"进贴图的光：一个物理直觉

在深入质检之前，先建立一个物理直觉，它能解释"为什么 AI 纹理总有点怪"。你看到的物体外观，从来不是一个纯粹的属性，而是三样东西相乘的结果：

> *观察到的颜色 ≈ 反照率（材质本来的颜色） × 光照与阴影（这一刻的光怎么照）*
>
> 反照率（albedo）是"物体自身反射多少光、什么颜色"的固有属性；光照则随环境变化。照片记录的是两者的乘积。

问题在于：一张照片记录的只是"乘积"，而我们要的材质是"被乘数"。同一个乘积有无穷多种拆法——"一块深灰的板被强光照亮"和"一块浅灰的板处于半暗中"可以给出完全一样的像素。AI 从图片里学材质，就像一个人只看过"穿上衣服的照片"去猜"本人的身材"：它可以猜得很好，也可能把阴影当成颜色、把高光当成图案，永久地"烤"进贴图里。研究者把这类问题归纳为"材质、光照与几何未被解耦"：模型的输入是观测值，而它需要的是成分值；更麻烦的是，它还会把训练图片所在渲染器的"风格偏好"一并学走——同一件 AI 生成的物体在 A 引擎里看着正常，换到 B 引擎里因为光照模型不同而显得"脏"或"糊"。

更糟的情形是：模型把"光照"当成了"物体的一部分"。生成一个"夕阳中的楼台"，它可能把暖橙色的光直接涂进青瓦的贴图里。你把模型摆到夜景的蓝光下，瓦片依然泛着夕阳色——因为那不是光，是它长在皮肤上的颜色。这就是为什么"材质要与光照解耦"是专业流程的一条铁律：**贴图负责"它是什么颜色"，引擎负责"光怎么照它"；两者混在一起，模型就失去了"能被重新打光"的能力。**

◆ 4.4.3 多视角一致性：背面幻觉的数据根因

现在回到本章开头小舟遇到的怪事：为什么模型正面像模像样，背面总是出事？这不是偶然，研究界已经把它追到了源头——**数据**。

有研究者用大规模图文数据集抽查后发现一个系统性偏差：互联网图片里，物体的正面被反复拍摄、大量出现；侧面特别是背面的图片则严重不足。想想你自己拍照的习惯就很明白——摆好姿势的是正面，风景里最常见的是你能看到的那些面。于是训练出来的模型自然"正面见得多、背面见不够"。这份偏斜的数据分布，会通过两条渠道放大到你的结果里：走 SDS 路线时，二维评委只在它见过的角度上可靠；走原生三维路线时，编码器同样把这份"正面丰盛、背面贫瘠"的先验带进了几何推断。研究者的结论毫不留情：现有各种缓解手段（多视角输入、人工修补）都停留在采样与推理环节，**没有任何方法改变了数据分布的根因。**

这个根因开出的"症状清单"，你应该能一眼认出：**第一**，背面几何糊成一团——不是没算，而是没有"见过"参照，只能泛泛而填；**第二**，背面长出多余结构——如把侧面的带子延续成"两条背带"；**第三**，也是最容易在质检时抓到的——背面出现了正面才该有的东西，社区案例里记录过"后视图里错误地出现人脸""背面的花纹是正面花纹的镜像复制"这类典型错误。研究者用形式化方法分析过单视图条件下的歧义，结论与 4.2 节一致：**信息不足的地方，模型只能"以偏概全"，而它见过的"全"本身也是偏的。**

同一份数据偏差还会在纹理上留痕：同一块布面在不同角度下颜色微微"流动"（研究者称之为色偏伪影），或者纹理在接缝处错位——因为不同视角的颜色本来就来自不同次推断，重投影时谁也无法说服谁。

◆ 4.4.4 质检方法：用"环绕截图"替代"转着看看"

专业流程不会"转着看看觉得还行"就算过。可复制的方法叫**多视角截图质检**：把物体固定，相机以相等的角度间隔环绕一圈，逐张截下关键视角，逐张检查。它之所以有效，是因为它把"我看看"变成了"逐项核对"——盯着 8 张定格的图找缺陷，比拖着鼠标转圈找缺陷，发现率高得多。而且截图能留档：你会发现缺陷、定位缺陷、在修补后回归验证，全靠这一组截图。

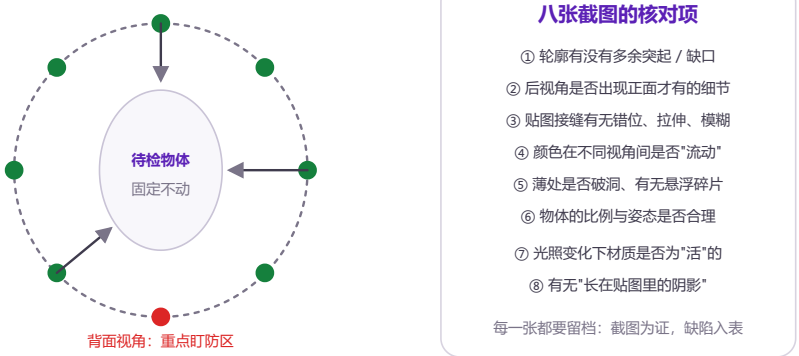


图 4-8　环绕截图质检。相机的 8 个等角位依次截图，其中"背面"是重点盯防区——那是模型最缺见识、也最容易胡编的方向。方法虽土，却是本章实践环节的第一件兵器。

为什么"背面重点盯防"值得写成纪律？因为一项大规模人评研究揭示了令人不安的事实：**我们（包括专业评测平台上的玩家）在评三维模型时，很容易被"渲染方式与贴图"带跑。**一份收集了 12.3 万余次投票、8,096 名用户、19 个模型的评测平台数据显示：同一批模型，用高斯泼溅渲染比用网格渲染平均高出 **16.6 个 ELO 分**；而"有贴图"比"没贴图"高出惊人的 **144.1 个 ELO 分。**研究者不客气地总结：人类在奖励渲染格式和贴图，而不是几何正确性。另一项缺陷评测里，研究者让 12 个视觉大模型充当"质检员"，最好的那个依然落后于经过训练的人类标注者。

这对你的直接启示有三条：**第一**，别用"带贴图的移动渲染"当验收标准——把贴图关掉、把线框打开，几何的真相才露出来；**第二**，在人评里"好看"与"正确"是两回事，答辩展示可以讲观感，交付验收必须讲几何；**第三**，既然 AI 质检员（视觉大模型）暂时顶不上人，**"环绕截图 + 缺陷记录表"这套人工流程就不是笨办法，而是当前唯一可靠的流程。**本书把"多视角质检记录"列入每个道具的交付物清单，理由就在这里。

表 4-6　多视角质检：逐项核对清单（建议打印或做成网页表单）

检查项	看哪里	常见缺陷	判定与动作
背面结构	后侧 1–3 张截图	糊成一团、长出多余结构、正反面细节错置	轻微→标注留档；严重→补多视角输入重生成
轮廓完整性	全部 8 张	缺口、多余突起、悬浮碎片	悬浮碎片必删（后续会飞进画面）
薄壁与镂空	薄板、檐口、栏杆处	破洞、双层壳、透过率不自然	破洞→补面；双层壳→交代为"已知缺陷"
贴图接缝	纹理分界处	错位、拉伸、模糊、色差	记录接缝位置，修补时优先处理
颜色一致性	相邻视角对比	同一表面颜色"流动"、明暗跳变	明显者重做贴图；轻微者记录为"观感风险"
材质活性	换光照再截一组	阴影与高光"焊"在贴图上、无法重新打光	判定为光照烘入，夜间场景慎用
比例与姿态	前侧对照截图	头身比失衡、姿态倾斜、轴线不水平	记录尺度数据，进引擎前校正

案例 4-A § 扁舟背面的一支"不该出现的桨"

背景：小舟用第 2 章的水墨分镜画面做图生三维，得到一艘扁舟。正面视角惊为天人：船身弧线、篷顶的墨色渐变都在。

做法：按流程绕圈截了 8 张图。前侧 6 张全过，第 7 张（后侧）出现异常：船尾多出一支笔直的木桨，而分镜里根本没有桨——那是模型"见过"的其他船留下的经验，被一股脑补进了没有任何观测支撑的背面；同时篷顶贴图在后视角整体偏灰，与正面墨色对不上。

结果：小月没有一键重生成整艘船（心疼已经调好的正面），而是把多出的桨剪掉、补了背后一小块篷顶的面，重投影了后侧的贴图。返工 40 分钟，船留下了。

复盘：这艘船的教学价值大于成品价值——它让小组明白：**背面缺陷不是"运气不好"，而是单视图输入的必然后果**；而且"局部修补 + 记录在案"往往比"重新抽一次卡"更划算。缺陷记录表上这一条被命名为"无观测来源的多余结构"。

4.5 拓扑、面数、UV 与 PBR：把模型"修成能用的样子"

从这一节开始，视角从"AI 怎么生成"切换到"你怎么接手"。生成模型交付给你的，是一件**看起来对的物体**；而工程链要的是一件**结构上对的资产**。这两者之间的差距，就是本节四个概念要补的课——它们也是行业面试里"懂不懂三维"的分水岭。请注意这四个词的共同点：它们在预览窗口里全都看不见。

◆ 4.5.1 拓扑：网格的"骨架排布"

拓扑（topology）说的是网格的"连接方式"——哪些顶点连着哪些面、面与面怎么排布。它不改变物体的外观（外形由顶点位置决定），却决定这件资产**能不能被后续使用**。

行业对"好拓扑"有几条硬标准，逐条讲：**①流形**（manifold）——一条边最多被两个面共享、没有悬空的边。非流形网格像一件"有裂缝的三合板"，3D 打印、布尔运算、物理模拟都会在裂缝处出问题。**②水密**（watertight）——网格完全封闭，像一只"不漏水的瓶子"。想打印它、想给它算体积、想让物理引擎判定"墙不可穿过"，都要水密。**③四边形为主**——由四个顶点组成的面（quad）在细分、雕刻、动画变形时的行为远好于三角形，所以绑定骨骼用的模型通常要"四边面化"。**④边循环**（edge loop）沿形变方向排布——肘关节、肩部这些要弯折的地方，面必须横着"圈"住关节，否则一弯就破。这一条在你做角色动画时（第 5、6 章）是生死线。**⑤没有多余的内部面、没有 n-gon**（超过四边的面）——它们会在打磨与导出环节制造意外。

来看看 AI 生成的网格为什么总在这几条上栽跟头。它们的几何多半来自"隐式场提取表面"的过程——用 marching cubes 之类的算法把一个连续场"切"成三角面。这类算法的产物有一个鲜明特征：成千上万个大小相近的三角形，像在表面撒了一层均匀的芝麻。研究者对 AI 网格的这句描述值得原文引用：**"这些网格通常带着数十万个近乎均匀的三角形——对渲染、编辑与动画管线而言都过于稠密"**。均匀三角面本身不是罪；罪在它**不沿形变方向排布**、不照顾关节、还常常带着孔洞与非流形边。同一份研究同时给出了让你安心的一半：研究界也承认"非水密、非流形的网格在真实生产管线里依然是有效且有用的资产"——**水密并非所有生产管线的统一门槛，而是取决于用途**。判断标准是场景：要打印、要物理，就水密；只是摆景、只当远景道具，允许通融。

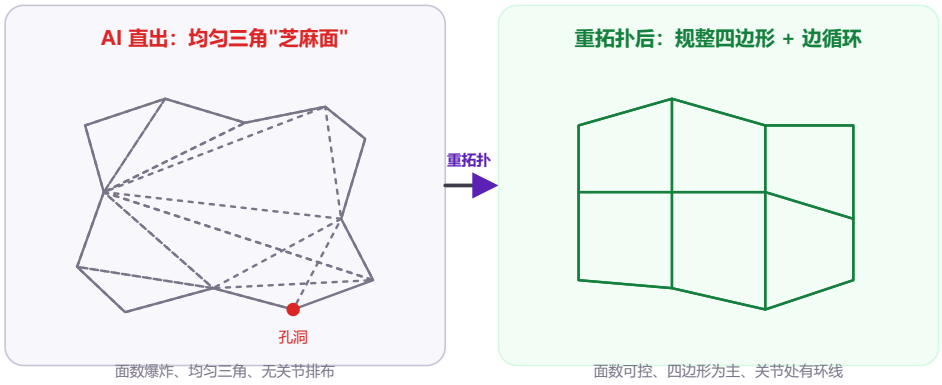


图 4-9　同一件物体，两种拓扑。左边是 AI 直出的"均匀三角芝麻面"（示意），右边是重拓扑后的四边形网格。外观几乎一样，可用性天差地别——左边不能绑骨骼、不能快乐地改，右边可以。

◆ 4.5.2 面数：一道预算题，不是一场军备竞赛

面数（poly count）是资产上三角形（或四边面折算成三角形）的总量。新手常犯的错误是"面数越多越好"——上了 AI 工具，看默认导出 20 万面，觉得赚了。真实世界里，面数是**预算**：场景的总算力、显存、加载时间都在预算里拨款，道具与道具之间是竞争关系。你给楼台多花 90% 的面数，等于从江水、灯笼、扁舟身上各砍一刀。

行业给出的经验区间（这些数字来自生产实践总结，可作为你们的起步档）：**移动端背景道具**约 200–1,500 个三角形；**桌面端环境道具**约 1,500–8,000；**主角级道具 / 武器**约 8,000–25,000；角色工作流的目标约 2 万面上下。而 AI 工具直出的网格常见在**5 万到 50 万面**之间——比"游戏就绪"的标准高出十几到几百倍。对照两个坐标你就懂了：Tripo 的 Smart Mesh 模式默认产出约 **5,000 面**的游戏就绪拓扑；而手工制作场景 vs AI 辅助场景的对照案例里（一个 n=1 的个别案例，不能当普遍规律），同一场景手工 123 件物体合计约 29.4 万三角形，AI 辅助 103 件物体合计约 **3,278 万三角形——约 111 倍**，也制造了后续的返工债。研究者对 AI 生成网格"过于稠密、不适合渲染编辑动画管线"的批评，数据上就是这么来的。

还有一个今年的重要变量：**UE5 的 Nanite** 这类虚拟几何技术大幅放宽了"面数"约束——几十万面的资产可以直出直用。但别高兴太早：放宽的是"渲染面数"，不是"结构要求"。拓扑依然要干净（否则 Nanite 聚类出错）、UV 依然要合理、贴图显存依然是更早撞上的天花板（在多数项目里，显存告急往往先于面数告急）。而在 UE5 之外的引擎、移动端、Web 端展示、3D 打印这些场景里，面数预算一分都没松。

面数预算还引出本章的一个新名词：**细节层级**（level of detail，LOD）。同一个道具通常要准备几个精度档：离镜头近时用高精度版本，中距离换中档，远处的路人与背景物件用最低档——引擎按距离自动切换，这就是"LOD 链"。为什么要提它？因为AI 工具导出的资产通常只有一档精度，而且缺少经过验证的 LOD 链（以及碰撞体，4.6 节细讲）。也就是说，"面数超标"只是第一层债，"没有分档方案"是第二层债——你不只要把面数压下来，还要为它规划"远看是什么样、近看是什么样"。这也是本章的实操检查表里"附加物"一栏永远存在的原因。

处理超标面数有两条路，必须分清：**减面**（decimation）是"机械删三角形"——快，但会破坏四边形、把排水口变成烂牙，只适合"再也不会被编辑"的成品。**重拓扑**（retopology）是"照原样重建一张好网格"——慢（一个中等道具手工约 30–60 分钟，一个角色约 2–4 小时），但得到的是能进管线、能绑骨、能交给下一环的资产。经验法则：**素材级/远景用减面，主角级/要被动画的资产用重拓扑。**不想全手工，可以用自动化工具起个头：QuadriFlow、Instant Meshes、ZRemesher 都能从稠密网格重新生成四边形为主的低模，随后人工收尾（哪里该加边循环、哪里该合并）。AI 工具自己也在补课——例如 Tripo 的 Smart Mesh 与 Meshy 的原生四边形重拓扑，目标就是"直出游戏就绪拓扑"。2026 年 1 月正式发布的 Meshy 6，把"原生四边形重拓扑"写进了产品能力，值得关注。

表 4-7　面数预算参考（起步档，来自生产实践总结）

资产类别	三角形数量级	说明
移动端背景道具	200–1,500	远景物件、可批量实例化；面数就是电量
桌面端环境道具	1,500–8,000	本书短片的默认档：楼台、扁舟、灯笼都落在这一档
主角级道具 / 武器	8,000–25,000	会被镜头怼脸看的资产
角色（含绑骨）	约 20,000 上下	需要边循环与干净拓扑，面数值可放宽、结构不可放宽
AI 工具直出	50,000–500,000	比就绪标准高十几到几百倍；参考：Tripo Smart Mesh 默认约 5,000 面

表 4-8　减面与重拓扑：什么时候用哪条路

方法	手段	耗时	产物质量	适用场景
减面（decimation）	算法直接删三角形（保留形状的粗糙近似）	秒级–分钟级	三角面、可能破四边结构、关节处不可用	远景、静态素材、不再编辑的成品
重拓扑（retopo）	照原件重新描出四边形网格（自动工具起头 + 人工收尾）	中等道具约 30–60 分钟；角色约 2–4 小时	四边形为主、有边循环、能绑骨、能进管线	主角资产、需要动画、需要长期维护

案例 4-C § 同一个场景，111 倍的面数：一次个案对照的启示

背景：本地研究报告记录了一次小规模对照实验：用同一美术方向制作同一个游戏场景，一组完全手工，一组让 AI 辅助生成各类物件。两组都被要求达到"可交付"的状态。

做法：清点两组的物件数量与三角形总量，并记录 AI 组后续的返工工时。

结果：手工组 123 件物体，合计约 **29.4 万**个三角形；AI 辅助组 103 件物体，合计约 **3,278 万**个三角形——总量相差约 **111 倍**，同时还积累了数百分钟（478 分钟）的返工债。

复盘：读这个案例要按住三件事：① 这是 **n=1 的个案研究**，不能推广成"AI 一定导致 111 倍膨胀"；② 它真正说明的是——AI 省下的是"从零建模"的工时，但如果不做减面/重拓扑，这部分成本并没有消失，只是转移成了"优化工时"和"运行时开销"；③ 因此"面数预算"必须写进你们的验收标准，而不是等场景卡了再补。它和 4.5.2 的预算表正好互为佐证：工具给你的是"未经协商的稠密几何"，而预算是你必须替项目把关的数字。

◆ 4.5.3 UV 展开：把三维的"皮肤"摊平成二维

贴图是一张平面图片，而模型是一层曲面——想让二者对上，必须先把曲面"剪开摊平"，建立"模型表面的每个点 ↔ 图片上的每个像素"的对应表。这份对应表就是 **UV 坐标**（U、V 是两个方向轴），这个裁剪摊平的过程叫 **UV 展开**（UV unwrapping）。

它像做一件玩具的"衣服纸样"：要把整张皮完美摊平是不可能的（球面无法不拉伸地展成平面，这是几何性质），所以必须在某些地方**剪开**。剪开的位置叫**接缝**（seam）。UV 展开的全部手艺都在这条接缝上：

- **接缝要藏在天然分界处。**房檐、墙角、木纹方向的变化处——观众不会盯着那里找破绽。把接缝剪在瓦片正中央，纹理断开就会刺眼。
- **不许重叠。**两个面共用同一块 UV 区域，贴图就会"互相涂抹"。AI 导出常见的 UV 混乱（重叠、压扁、极端拉伸）多源于自动展开工具在复杂拓扑上的失效——而复杂拓扑，正是 AI 直出的常态。这也是"**先重拓扑、再展 UV**"这条次序铁律的由来：在混乱拓扑上展 UV，等于在打结的毛线团上绣花。
- **留出边距。**不同的 UV 块之间要留一点空隙（常见做法是几像素的 padding），否则贴图缩小时颜色会互相渗透。

- **纹素密度要匹配。**纹素密度（texel density）指"模型上一个单位长度分摊多少贴图像素"，行业常用约 512 像素/米作为基准。同一场景里的物件密度差太大，就会出现"这件糊、那件锐"的穿帮——观众说不出哪里错，但能感觉到"不干净"。

AI 生成资产时，UV 往往是"顺带送"的：几万几十万面的拓扑直接自动展开，重叠与拉伸在所难免，贴图错位、糊斑通常都能在这里找到病根。修补时的正确顺序是：**重拓扑 → 重新展 UV → 重投影或重画贴图**——反过来做全部白费。

◆ 4.5.4 法线与烘焙：让低模穿上"高模的外套"

低模面少、轮廓糙，高模面多、细节足——能不能"看起来像高模、跑起来是低模"？可以，这就是**法线贴图**（normal map）的功劳。它的原理可以这样理解：模型表面的每个点都有一个朝向（法线方向），光照的明暗取决于这个朝向。法线贴图不改几何，只改每个点"看上去的朝向"——于是平面的地方能"画"出凹凸、螺丝能"长"在皮肤里，渲染成本几乎为零。

把高模细节"转移到"低模贴图上的工序叫**烘焙**（baking）：把低模与高模对齐叠放，逐个像素问"在这一点上，高模的法线朝向是什么"，写进贴图即可。烘焙也是"把程序化材质、光照信息、材质分区固化进贴图"的通用手法。这套技术在 AI 时代反而更重要了：AI 生成的模型面数巨大但结构混乱，专业流程常常是"把 AI 结果当高模，重拓扑出低模，再烘焙回细节"——AI 负责提供"像什么"，你负责提供"能用什么"。

◆ 4.5.5 PBR：材质不是"一个颜色"，是"一组通道"

PBR（physically based rendering，基于物理的渲染）是当代材质的通用语言。它的核心思想只有一句：**用物理上讲得通的参数描述材质，让同一份材质在任何光照环境下都有正确表现。**落到文件上，就是一组各管一件事的"通道"贴图。必须记住每个通道"是什么、谁来管、常见的错法"——这是导出环节 90% 事故的策源地。

表 4-9　PBR 通道：各管什么，常见错法是什么

通道	管什么	取值直觉	常见错法
反照率 / 基础色 albedo / base color	材质"本体"的颜色	纯"颜色"，不含光效	把阴影、高光、环境光"烤"进颜色；文字与栏杆的对比度里混入光照
法线 normal	每个像素"看上去"的朝向	紫蓝色的方向编码图	被当普通彩图处理（做了伽马校正）→ 光影全错
粗糙度 roughness	表面多"哑"	0 = 镜面；1 = 全哑光	与光泽度（glossiness / smoothness）混用而不取反
金属度 metallic	材质是金属还是非金属	1 = 金属；0 = 绝缘体（木、石、布）	取值"半金属"造成物理上不存在的材质
环境光遮蔽 AO	凹陷处的柔和暗角	灰度图，越黑越"被挡"	当阴影用、叠得过重导致"脏"
高度 / 置换 height / displacement	表面的起伏信息	灰度图，亮 = 高	开了置换却不加细分，表面"炸刺"

这组通道多数遵循"金属-粗糙度工作流"（metallic-roughness）：把金属度与粗糙度打包进同一张图的不同通道，节省显存。打包方式在各大引擎里各不相同——这正是"导出事故"的高发区。请认真读表 4-10：同一个模型，在 Blender 里正常、到 Unity 里发粉、到 Unreal 里发灰，原因常常不是模型错了，而是通道的"装箱方式"与"色彩空间"约定没对上。

表 4-10　同一份 PBR 数据，各家的"装箱方式"并不相同

环境	粗糙度 / 光泽度	通道打包方式	要小心的坑
Blender（Principled BSDF）	直接有粗糙度输入	各通道分开接入，概念最直白	导出时按目标引擎打包，别把 Blender 的约定当通用
Unity URP Lit	暴露"光泽度 smoothness" = 1−粗糙度	金属度图：金属度 R 通道 / 平滑度 A 通道	从别的软件搬材质要先"取反"；不取反 → 粗糙物反光、金属发灰
Unity HDRP	掩码图 Mask Map 统管	金属度 R / AO G / 细节 B / 平滑度 A	通道塞错位置，症状是"能渲染但全不对"
Unreal Engine	独立连接粗糙度	通道分开；数据类贴图要关闭 sRGB	法线/粗糙度/金属度被当色彩图做伽马校正
glTF / Three.js（Web 端）	粗糙度在图里	金属度 → 蓝色通道；粗糙度 → 绿色通道；AO 独立一张图取红色通道	手改通道顺序导致错位；Unity → glTF 的转换要注意平滑度取反

把这一节压成口诀再走：**拓扑决定"能不能用"，面数决定"在预算里能不能用"，UV 决定"皮肤贴不贴得住"，PBR 决定"换到哪儿都不塌"。**四件事在预览窗口里全都看不见，全部靠线框视图、UV 视图与平面贴图来找。**到 4.9 节的实践里，你会第一次亲手把它们一件件查出来——那时你才会真正明白，为什么"出图好看"和"资产合格"之间隔着一整条产线。**

4.6 导出、交接与检查清单：让资产"能出货"

模型看顺眼了，缺口补了，面数也压了——最后一步是把文件"交出去"。老手都会告诉你一句话：“引擎导入决定资产能否出货。”注意这句话的狠处：决定资产命运的，不是预览窗口里它多好看，而是它在 Unity 或 Unreal 里被拖进去那一刻发生什么。一次失败的交接，能让前面所有的工作一秒作废。

好消息是，交接事故高度可预防。事实材料里有一条非常具体的经验结论：**一张“导出前检查表”能捕获约 90% 会破坏 Unity / Unreal 导入的问题。**这一节把这张检查表拆开讲清楚。先讲清楚"导出"到底在发生什么——你会发现，导出从来不是"保存文件"，而是**把一种约定翻译成另一种约定**；事故都发生在翻译的缝隙里。

◆ 4.6.1 格式：三种文件各司其职

你会在导出口看到一堆后缀名。不用全记，但要记住三种"主力"和它们的性格：

- **OBJ**：最老牌、最通用的几何交换格式。什么软件都认它，适合"把几何搬来搬去"；但它对 PBR 材质、动画、骨骼的支持都很弱——它不知道什么是"粗糙度"，也不会带你的骨骼动画。
- **FBX**：Autodesk 家族的格式，**引擎工作流的事实标准。**游戏引擎的资产管线十有八九绕着它转；它能装下网格、材质引用、骨骼与动画。缺点：是私有生态的产物，跨工具时偶有版本脾气。
- **glTF / GLB**：Khronos 组织维护的开放标准，专为“实时渲染传输”设计，对 PBR 材质原生友好，Web 端（如 Three.js）的默认选择。`.glb` 是它的二进制单文件版本——推荐把它当成**跨工具交换的"友好默认"**：讨论、评审、网页预览用它，最大概率“别人也打得开”。

此外还有两个经常出现但不是"主力交接"的格式：**STL**——3D 打印的行业通用（只有几何、没有颜色贴图，而且对水密要求苛刻）；**USD / USDZ** 与 **3MF**——前者来自影视与工业管线（AR 预览常见 USDZ），后者是带颜色的打印格式。看清一个工具的导出列表，你就能反推它想服务谁：一个只给 STL 的工具，是给打印机用的；一个给全套 OBJ/FBX/GLB/USD/USDZ/STL/3MF 的（比如 Tripo 这类导出格式最全的工具），是想同时服务游戏、工业与打印。

表 4-11　常用导出格式的性格对照

格式	装得下什么	谁在用	什么时候选它
OBJ	几何 + 简单材质引用	建模软件、雕刻软件之间搬几何	只要几何、跨软件兼容优先；不指望 PBR 与动画
FBX	网格 + 材质 + 骨骼 + 动画	Unity / Unreal 的资产管线	进游戏引擎的"正经资产"，尤其要绑骨骼动画的
glTF / GLB	PBR 材质 + 网格（可含动画）	Web 端（Three.js 等）、跨工具预览	讨论评审、网页展示、跨工具的"友好默认"
STL	只有几何，无水密即失效	3D 打印（切片软件）	要打印时；先过水密检查再导出
USD / USDZ • 3MF	场景描述 / 带色打印	影视工业管线 / AR / 彩色打印	进入更大的工业管线，或带颜色的打印

◆ 4.6.2 尺度、轴向与原点：三个不写在模型上、却最常炸掉的坑

这一小节的三件事，是"引擎导入陷阱"里的前三名，也是新手最容易忽略的——因为它们在预览窗口里统统看不出来。你看到的模型永远"站在原点、大小正好"，可一旦跨过软件边界，约定就变了：

① 尺度（单位）。不同软件对“1 个单位”的定义不同：Blender 与 Unity 习惯 1 单位 = 1 米，而 Unreal 的默认单位是厘米。一条 1.8 米高的横杆，用米制写成“1.8”，进了按厘米理解的软件就变成“1.8 厘米”——或者反过来，被放大 100 倍。事实材料把它列为“最常见的失败点”，并给出典型症状：**100 倍的单位错配。**小舟那个“楼台比桥大近百倍”的怪事，多半不是 AI 手滑，而是这 100 倍在交接时发生了。

② 轴向。三维世界有三根轴，但各家对“哪根朝上”的约定不一样：Blender 习惯 **Z 轴朝上**，而 Unity / Unreal 习惯 **Y 轴朝上。**差这一次旋转，模型进场时会**"横着到达"**——房子躺着、小人趴着。这正是开头那个“楼台躺在地上”的病因。

③ 变换与原点。如果你在建模软件里把模型移动、旋转、缩放过，但没有执行“应用变换”，导出的文件里就会带着一份“待价而沽的缩放数据”。事实材料的原话很生动：未应用的变换 **“到达 Unreal 会变成一个煎饼”**——因为引擎按自己的方式解读了那份残留的缩放。同时，模型的**原点**（pivot，"旋转与摆放的把手"）该在哪，行业有约定：环境物件放**底面中心**（拖进场景就往地上放）、角色放在**颈根/下巴附近**（便于抬头低头与绑骨）。原点偏在角落里，就像"门把手长在门缝边"，摆放成为噩梦。

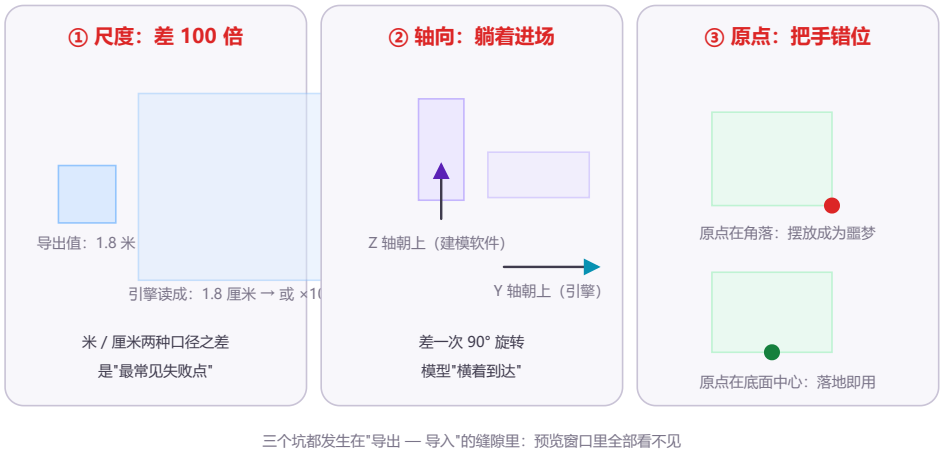


图 4-10　交接的三大陷阱：尺度、轴向、原点。它们不改变模型本身，却能决定模型在目标软件里"是否还能用"。导出检查清单的前三条，逐条对应这三格。

◆ 4.6.3 “确定性断言”：截图证明不了什么

你可能想说：我导入之后截个图，看起来对，不就行了？来学一个新词：**确定性断言**（deterministic assertion）——指的是“可以用程序判定真假”的检查，比如“缩放值是否等于 (1,1,1)”“非流形边的数量是否为零”“导出再导入一圈后顶点数与贴图是否一致”。与之相对的是"外观检查"：转盘截图、环绕预览——它能证明的只有**画面看起来如何。**

为什么非要较真？因为三维资产的大部分真实缺陷是"看不见"的：一个 100 倍缩放在小部件场景里看不出来（周围没有参照物），一次轴向旋转在对称的模型上可能看不出来，一个非流形边在光滑着色下根本看不出来——它们只在被使用的那一刻爆炸：绑骨骼时炸、物理碰撞时炸、打印切片时炸、灯光烘焙时炸。所以专业流程把"断言"前置：能写程序查的，绝不交给眼睛。事实材料还记录了一条更扎心的提醒：**AI 工具导出的资产通常缺少经过验证的碰撞网格与 LOD 链**——这两样东西恰恰是"进引擎能不能用"的硬指标。也就是说，AI 给你的是一"件外观"，不是一"份资产"；从外观到资产之间的差价，要你自己补。

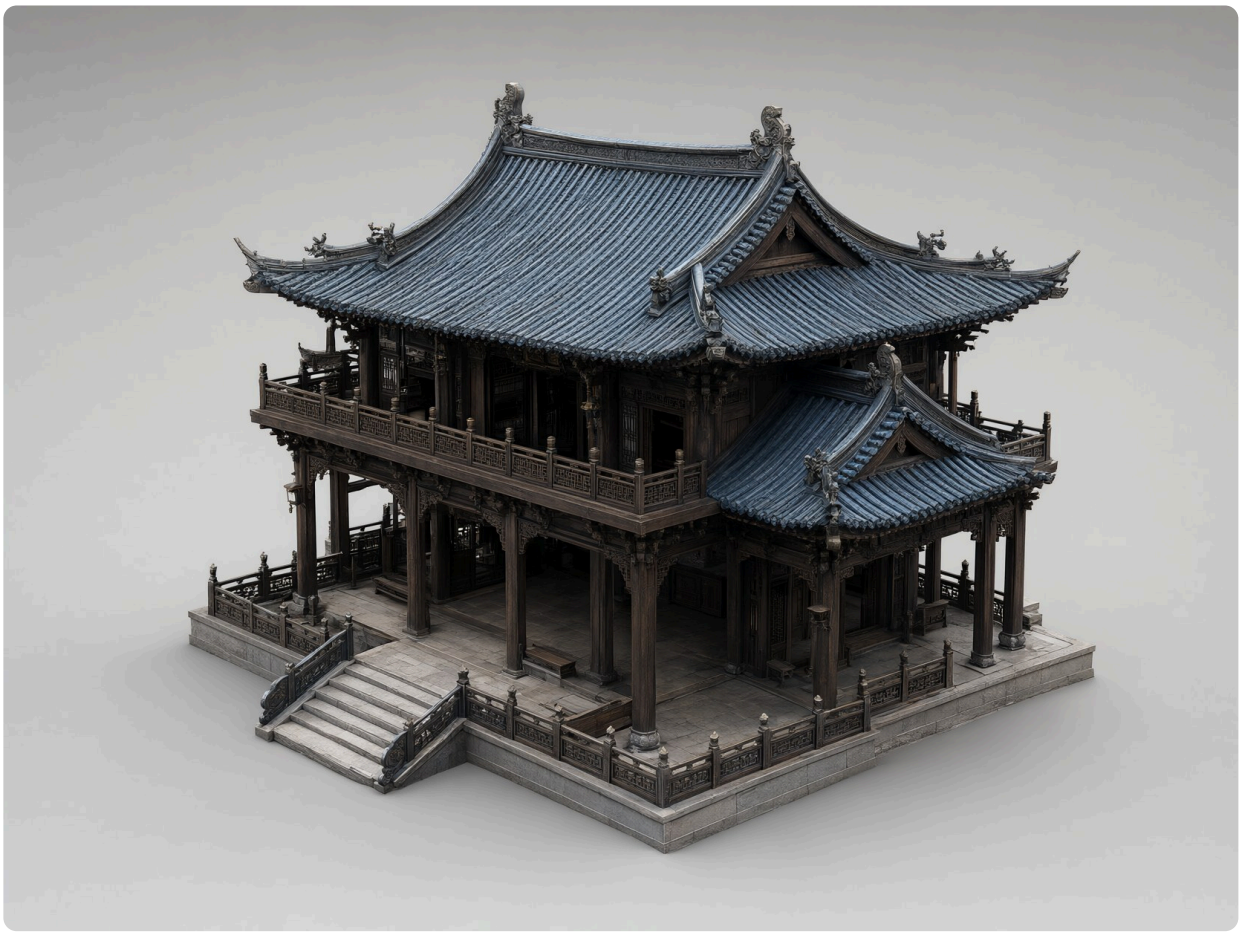


图 4-11　一个「可以直接送检」的资产长什么样。中性灰底、正交视角、无影棚拍光——这是三维资产预览的标准呈现方式：去掉了环境干扰，任何结构缺陷（缺角、破面、法线翻转）都无处藏身。本章的「环绕截图质检」请按这个标准拍。 AI 生成

◆ 4.6.4 导出检查清单

把前面的知识合订成一张可以打印出来的表。这张表按"翻车频率"排序：前几条几乎人人中招，后几条是质量加分。养成习惯：每一项都必须有"证据"（一个数字、一张截图、一次断言结果），不许凭记忆打勾。

表 4-12　导出检查清单（覆盖约 90% 的导入事故）

#	检查项	具体动作	合格标准（要有证据）
1	**尺度**	量出真实尺寸（如楼台高 6 米），核对导出单位与目标软件单位	在目标软件里量出的尺寸与预期一致（±数厘米）
2	**轴向**	确认“哪根轴朝上”的约定，必要时旋转校正	导入后模型站立而非躺着；地面接触面正确
3	**变换与原点**	应用全部变换（缩放归 1）；原点放到约定位置	缩放值 = (1,1,1)；环境物件原点在底面中心
4	**法线朝向**	检查背面剔除下是否有黑面/空洞	外表面全部朝外；无内外面反转
5	**拓扑与面数**	线框视图数一遍；核对面数预算	无多余内部面、无悬浮碎片；面数在预算档内
6	**水密（若下游要打印/物理）**	用切片器或检查工具跑一遍	无开放边、无非流形边、无自相交
7	**UV 与贴图**	UV 视图检查重叠与拉伸；贴图分辨率与纹素密度	无重叠、无明显拉伸、贴图随文件带走
8	**通道与色彩空间**	数据类贴图（法线/粗糙度/金属度）关闭 sRGB	目标软件里明暗正确；Unity 不出现"粉红材质"
9	**命名与组织**	文件名、物体名、材质名统一规范	不出现"未命名 27""材质 03"这类名字
10	**附加物**	清点碰撞体与 LOD 链是否存在、是否验证过	缺则明确标注"缺失"，列入下游工作
11	**导出设置**	勾选目标格式选项；确认贴图"内嵌 / 随包"	对方不开你电脑也能打开、贴图不丢
12	**回导入验证**	把导出的文件重新导回目标软件，重跑断言	round-trip 后数值一致——这是最终裁决

注意第 12 条：**回导入验证（round-trip check）是把关的最后一关**——你把文件导出去，再把它导回一个干净的目标环境，重新量尺寸、数面数、看贴图、试摆放。事实材料称之为"导出往返测试"，它之所以是"最终裁决"，是因为它模拟的正是下游选手的真实动作。**自己导出的东西自己没回过一次门，就等于没检查。**

◆ 4.6.5 如果下一环是"绑骨骼"：一份特别的交接检查

第 5 章你要给角色做动画，绑骨骼之前的资产质量决定了一切后续。事实材料给了一份"经验清单"（再次强调：这是生产经验总结，不是强制标准）——进入绑骨流程前，角色类资产最好满足：

表 4-13　绑骨骼前的资产检查（经验清单，非强制标准）

检查项	为什么
标准姿态：T-pose 或 A-pose	绑定与动捕数据的通用参照姿态；某些工具会强制输出该姿态正是为了这一天
左右对称	让"镜像权重"可以工作；不对称的资产权重得手工两遍
干净网格：无内部面、无悬浮碎片	蒙皮算法会被多余几何骗，权重涂到看不见的地方
关节处的边循环	肘、肩、膝要能弯；没有环线一转就破洞
完整的 UV	贴图是"皮肤"，皮肤破了动画一样破相
单位与比例：1 单位 = 1 米（按项目约定）	动画数据、物理参数、引擎工具都按单位算
应用变换：缩放读出 (1,1,1)	带缩放绑骨 = 把弹性形变烤进动画里，后患无穷
法线朝外	渲染与烘焙的前提
面数在 LOD 预算内	骨骼是给"预算内资产"服务的；超标资产先减/重拓扑
无 n-gon（超过四边的面）	变形时 n-gon 的三角化方案不确定，动画会"抖动"
原点位置合理：环境物件放地面、角色放颈部	摆放与动画控制点的约定俗成

对照这份清单，你会明白为什么"AI 直出的角色几乎不可能直接绑骨"：面数超、拓扑乱、UV 糊、原点偏——四条全中。业界的应对是两头一起走：一头是"人重做"，另一头是"工具补课"——例如有的工具开始提供一键自动绑骨（Tripo 的 UniRig 宣称可以把骨骼一键装好、骨架与通用动作库兼容），有的生成模型干脆强制输出 T/A-pose，把"标准姿态"这一步做成默认。**但骨架装好不等于权重就好——自动绑骨的结果依然要人工检查与喷涂，这一步第 5 章会详细讲。**

◆ 4.6.6 引擎导入常见陷阱速查

把"交接缝隙"里的全部高发事故汇总成最后一张速查表。建议你们组的"导出负责人"把这张表贴在显示器上——事故现场它比教科书有用。

表 4-14　引擎导入常见陷阱速查（按翻车频率排序）

陷阱	典型症状	根因	处方
尺度错配	模型大 100 倍或小 100 倍，"人和楼一样高"	米 / 厘米两种单位口径混用	导出前量尺寸、对单位；清单第 1 条
轴向错位	模型躺着进场、朝向与预期相反	Z-up（建模软件）与 Y-up（引擎）之差不处理	导出时校正轴向；导入后立即目检
变换未应用	进场后被压扁、缩放诡异（"变成煎饼"）	建模时的缩放/旋转残留在变换数据里	应用全部变换，确认缩放 (1,1,1)
原点错位	拖进场景后"人不在地面上"、缩放与旋转中心诡异	原点不在约定位置	环境物件放底面中心；角色放颈根
面数与拓扑超标	场景卡顿；细分/绑骨工具报错	AI 直出资产面数超预算十几到几百倍	减面或重拓扑到预算档
UV 混乱	贴图涂抹、错位、大面积糊斑	自动展开在混乱拓扑上失效	重拓扑 → 重展开 → 重投影
材质通道 / 色彩空间错误	金属发粉、粗糙物发亮、整体发灰	通道打包约定与 sRGB 设置没对齐	按目标软件重置通道；数据贴图关 sRGB
"粉红材质"（Unity 高发）	整个模型粉红色	材质 / 着色器未能正确交接（管线不匹配）	换成目标渲染管线的材质；重指定一次
缺碰撞体 / LOD 链	物体可穿模；远处物件依然吃满算力	AI 导出不含这两样，且没有验证过	明确标注缺失，列入下游工作清单
贴图未随行	对方打开后模型是纯灰的	导出了几何、没导出或没内嵌贴图	用带贴图的格式（如 GLB 内嵌）；回导入验证

最后补一句 3D 打印方向的下游差异，因为你可能选打印作业：打印对模型的要求比渲染**更苛刻**——它关心水密、壁厚与自相交（表面自己穿过自己）。事实材料给出的参考是：家用 FDM 打印机的可打印最小壁厚约 **1–1.2 毫米**，光固化（SLA）约 **0.6–0.8 毫米**——比这薄的壁，生成得再“好看”也会在切片时被删掉；而自相交几何占比超过约 **5%** 的模型，打印失败率显著更高（材料里给出的数字是约高出 40%）。同组数据还提到：同一批生成模型中，能宣称"完全水密"的比例并不高，最好的一家也谈不上普遍可靠，而部分工具的产物在水密率上明显垫底，所以——水密不是"默认就有"，而是"逐件验证"的。切片软件（slicer）因此成为打印方向的"引擎导入验证"：先切片，看它报什么错，再决定修什么。**4.8 节的失败诊所 F4-3 会给你看一个真实的切片灾难现场。**

4.7 工具地图：这一排名字，各管一段路

原理讲完，来认一认"实战地图"上的城市。三维生成的工具格局在 2025–2026 年变化极快——本节的信息都标注了口径与时间（具体功能与免费额度请以官方页面为准，这是本章反复强调的信息素养）。你可以把这一节当"选型备忘"：不需要背名字，需要记住的是**四类工具的分工**：云端生成服务、开源本地模型、场景级世界生成、以及把它们缝起来的专业软件。

◆ 4.7.1 云端生成服务：提交文字或图片，拿回模型

课程实践将主要在云端服务上完成，因为它们免安装、有免费额度、上手最快。四家主流服务的分工与要点：

表 4-15　云端三维生成服务选型备忘（数据取自公开材料，功能与额度以官方页面为准）

服务	来历与路线	面向谁	关键参数（公开口径）	注意
混元 3D 3.1（腾讯）	2026 年 1 月末发布；云服务	想"开箱即用"的个人与团队	每日 20 次免费生成；云 API 有 200 免费积分；默认导出 GLB；面数可配置约 5 万–150 万；提供"水密"开关参数	许可条款有地域与规模限制（详见 4.11）
Tripo	Algorithm 3.1 宣称超 2000 亿参数	游戏与管线型用户	导出格式最全（OBJ/FBX/GLB/GLTF/USD/USDZ/STL/3MF）；最高约 200 万多边形；8K 贴图；Turbo 模式约 10 秒；Smart Mesh 约 2 秒给游戏就绪拓扑；UniRig 一键自动绑骨	免费额度各来源说法不一（每月 200 或 300 积分，存疑）；免费层输出按 CC BY 4.0 署名；Pro 约 19.9 美元/月
Rodin（Gen-2 / Gen-2.5）	Gen-2 为 10B 参数级；Gen-2.5 于 2026 年 5 月亮相	游戏与工业设计	免费生成、付费下载；千万级多边形上限；4K PBR；强制输出 T/A-pose	STL 导出通常需要修复；不适合直接打印
Meshy 6	2026 年 1 月中旬正式发布	3D 打印与游戏双线	面数 5,000–100,000（默认约 3 万）；原生四边形重拓扑；水密性宣称约 97% 通过切片器	免费每月 100 积分（约 10 次下载），CC BY 4.0 需署名

另外两条“存疑线索”也值得记录下来作为信息素养的练习：有材料提到通义万相支持 OBJ / GLB 导出，也提到即梦有 Maya / Blender 插件与"三维白模预览"能力——这两条在本课程编写时**未能从官方渠道交叉验证，属于存疑信息**，使用前请以官方页面为准。为什么把"存疑"也写进教材？因为这就是真实的行业状态：三维工具的公开资料更新极快、转述失真率高，合格的创作者必须养成"三源核验"的习惯——官方文档优先，其次官方页面变动记录，最后才是二手文章；转述里的免费额度、参数上限、许可条款，一律当作"可能过期的传闻"。

◆ 4.7.2 开源与本地部署：把模型请回自己的电脑

如果你想要可控性——不依赖别人的服务器、不被额度限制、能改代码——开源方案是另一条路。四条值得知道的线：

- **混元 3D 2.1（开源版）**：2025 年 6 月中旬开源，两段式结构——形状模型 3.3B 参数（约需 10GB 显存）、贴图模型 2B 参数（约需 21GB 显存）。在 HuggingFace 上的下载量以百万计（公开材料称超过 180 万次），社区生态活跃。但注意许可：它用的是腾讯社区许可，并非 MIT；条款对商用规模、地域与"用输出改进其他 AI 模型"都有限制——4.11 节细讲。
- **TRELLIS.2**：4.3 节的主角，采用 MIT 许可（商用友好），输出带 PBR 的 GLB；门槛是需要 NVIDIA 显卡且显存不低于 24GB。适合"组里有强显卡 + 想深度定制"的路线。
- **混元 3D PolyGen 1.5**：做的是"原生四边形网格生成"——直接生成游戏管线喜欢的四边形拓扑，而不是先三角面再重拓扑。公开材料称它把过去"以周计"的重拓扑工作压缩到"以分钟计"。这条线值得盯：如果重拓扑被模型自动化吃掉，4.5 节里"重拓扑 30–60 分钟"的成本结构会被改写。
- **HY-World 2.0**：2026 年 4 月中旬开源的世界生成工具，能导出 Mesh / 3DGS / 点云三种形态，并支持 3DGS 与网格的混合——它是"表征融合"趋势在工具层的落地：背景要照片级真实（3DGS），道具要能交互（网格），一个场景里两者并存。

◆ 4.7.3 行业里的真实效率数字：好看，但要学会读

最后看一段行业案例。它们经常出现在新闻与宣传里，值得了解，但你必须学会给数字"验明正身"。

🩺 **案例 4-B** § 游戏大厂的三维资产产线改造

背景：多家国内游戏公司近年把 AI 三维生成接进资产产线。公开材料提到：腾讯的 VISVISE 已在近 100 个游戏项目中应用，覆盖《和平精英》《王者荣耀》这类头部项目，部分环节效率提升最高约 8 倍；某类三维资产的生产从"5–7 天、数千元"压缩到"10 分钟、数元"；轻量游戏的整体开发周期从约 1 个月缩短到约 2 天；一个 25 平方公里的虚拟城市搭建从 5 天缩短到 25 分钟。

做法：不是"一键生成整款游戏"，而是把 AI 嵌进既有管线：程序化生成地形与建筑体块、AI 生成道具初稿、人工负责精修与验收——AI 管"量"，人管"质"。

结果：材料还记录了盛趣游戏的例子：做《黑神话：悟空》的实景复刻时，团队只有 1 名三维美术 + 1 名场景设计 + 半个技术支持；后续项目《地城之光》的效率较手工提升约 100%。Meshy 官方也宣称"过去需数周、花费约 1000 美元的 3D 资产，现在约 2 分钟、1 美元"。

复盘：读这些数字必须加三个注脚：① 多数为**企业自报或券商估算**，未经独立审计；② "最高 8 倍""提升 100%"指的是**部分环节**，不是全流程；③ 所有案例的共同点是"AI 出初稿、人工精修"，与本章反复强调的质检与交接完全一致。**把宣传读成"造一个资产只要两分钟"的人，会在验收线上付出真实的学费。**

这份地图到手的用法是"反向推"：从你的需求出发问五个问题——① 我要不要绑定骨骼或做动画？（要 → 看拓扑与姿态能力，如 Smart Mesh / UniRig / 强制 T-pose）② 我要不要打印？（要 → 看水密与壁厚，STL 前先切片验证）③ 我手上有几张参考图？（多 → 优先支持多视角输入的工具）④ 我是交作业还是商用？（商用 → 逐条读许可，颜色最浅的 MIT 与"CC BY 4.0 需署名"最省心）⑤ 我电脑什么配置？（显存低于 24GB → 本地跑大模型基本出局，用云端；反之可以考虑开源路线）。工具会过时，这五个问题不会——它们是你在未来任何一波新工具里都能用的"选型算法"。

4.8 失败诊所：三类高发缺陷的完整病历

进诊所之前，先立一块"预期管理"的牌子。行业里有从业者做过粗略的经验统计：**大约每 10 次三维生成，只有约 1 次可以不返工直接进入使用**；据此给出的选型建议是——AI 三维生成适合做**背景道具、LOD 层级、简单的单体几何**，不适合主角资产，更不适合任何"需要被读出来的文字"。记住这句话的目的不是劝退，而是校准：它不是"工具很烂"，而是"工具是初稿机，不是终点站"。初稿机产出初稿，检验与精修是你工作的一部分——下面三份病历，分别是这门课里你最容易遇到的三类"初稿病"。

🩺 **失败诊所 F4-1** 楼台背面的一团浆糊：背面幻觉

症状：小舟的文字生成楼台，正面八分像：翘角、木格窗、栏杆俱在。转到背面——墙体塌成一团起伏的疙瘩，屋檐与柱子在某处"长"在了一起；更诡异的是，背面墙上出现了一段与正面一模一样的格窗图案，而按结构那里应该是素墙。

成因：这不是软件故障，而是**数据分布根因**。研究者抽查大规模图文数据后发现：互联网图片里物体的正面被反复拍摄、严重过采样，侧面与背面却严重不足；这份偏斜会经 SDS 或前馈编码器一路放大到几何推断里。另一条独立研究线用形式化方法证明了单视图条件的歧义性——给定一张正面图，背面的几何在数学上就是欠定的；案例库里甚至记录过"给一张正面的肩包图，模型在背面生成出两条背带"这类错误。还有研究给出了一个反直觉的警告：在固定视角数量的条件下，**把视角的跨度加宽，生成质量反而会下降**——因为跨度越大，越多的区域落在观测支撑不到的地方，模型只能编。

处方：① 输入侧：能多给视角就多给（正面 + 侧面 + 背面的三视图远胜单张图）；② 推理侧：部分流程支持注入视觉语言模型的"常识"来改善背面结构——有研究把背面几何的 Chamfer 一致性最多提高了约 37.2%，但请记住研究者的结论：**所有现有缓解都在采样与推理环节，没有任何方法改变了数据分布的根因**；③ 人工侧：按 4.4 节的环绕截图流程逐张验收，"背面"永远重点盯防；④ 修复侧：局部补面 + 重投影，通常比整件重生成划算（案例 4-A 的做法）。

失败诊所 F4-2 玻璃盏与船篷：薄壳与透明物体为什么最容易崩

症状：小月想给夜江场景做一只"玻璃盏"，输入提示词后拿到的模型：盏壁破了两处洞，转过去发现壁是"双层"的（像内外各贴了一层皮），透明度也不对——明明是玻璃，看起来却像磨砂塑料。另一路，扁舟的船篷（一片薄木顶棚）生成后同样出现双层壳与破洞，像一片"双皮奶"。

成因：两个叠加的原因。**第一，表示与算法的天生短板：**基于场式表示加表面提取的路线，在处理**开放曲面**（没有封闭体积的薄片）时经常出问题——研究界的描述是"往往会在开放表面上失败，产出薄薄的双层壳"。而透明与反光物体本身就是三维重建与生成的老大难：它们的光学表现不满足"同一个表面点在任何视角下颜色一致"的假设。**第二，提示词自己"招灾"：**工具的官方帮助文档明确提醒过——暗示透明或镂空的提示词（比如"玻璃罐""柳条篮"）得到的网格往往需要更多修复；文档给出的替代方案很具体：先把提示词改写成"陶瓷罐"，生成一个实体版本，再用材质与后续处理去表现想要的效果。

处方：① 改写提示词：用实体材质词替代透明/镂空词（"玻璃盏"→"青瓷盏"；"竹篮"→"竹编箱"），透明感留给材质与渲染去演；② 真要薄壳（船篷、纸灯），生成后用加厚工具补厚度，或改为"薄但封闭"的设计；③ 打印场景下，壁厚不低于打印机极限——家用 FDM 约需 1–1.2 毫米，光固化 SLA 约需 0.6–0.8 毫米，低于此值会在切片时被整片删掉；④ 验收时对"薄"的地方单独截图留证。

失败诊所 F4-3 五十万面的水杯：面数膨胀、拓扑混乱与水密失守

症状：小舟生成一只"白玉水杯"，导出时发现 50 多万面。拖进引擎：整个场景开始卡；拖进切片软件准备打印：软件劈头报错——先标出大量非流形边，接着暴露开放边，再检查又会发现部分面片法线翻转；最后自动修复把"太薄"的杯壁直接删掉，杯子漏了。

成因：三件事一起发作。**① 面数膨胀：**生成模型被训练成"优先外观、不顾效率"，产物常见数十万个近乎均匀的三角面——研究者对这类网格的评价原文是"过于稠密，对渲染、编辑与动画管线而言都不可用"；**② 拓扑混乱：**从场里"切"表面的过程（marching cubes 一类算法）产出的是均匀三角网格，天然带着孔洞、非流形结构，破坏了打印与物理所需的水密前提；**③ 水密率不稳定：**同批生成的模型里，能"完全水密"的比例并不让人放心——部分工具的水密率明显偏低；而自相交几何超过约 5% 的模型，打印失败率会显著更高（材料给出的数字约高出 40%）。

处方：① 生成时选对面数档（工具普遍支持配置面数上限，甚至提供"水密"开关）；② 下载后按用途分流——进引擎的走重拓扑/减面，打印的先过切片器验证；③ 切片器（或水密检查工具）是打印方向的"回导入验证"，它报什么先修什么：非流形 → 开放边 → 法线 → 壁厚，按序处理；④ 验收断言：面数（数量）、非流形边（应为 0）、自相交（比例）三个数字必须落在记录表里。

表 4-16　失败模式速查表（症状 → 成因 → 处方）

失败模式	典型症状	根因一句话	首选处方
背面幻觉	背面糊、长出多余结构、后视图出现正面细节	数据里缺乏背面样品，单视图信息欠定	多视角输入 + 背面重点质检 + 局部修补
多面神（Janus）	转一圈发现多个"正面"（多张脸/多个同款装饰）	二维监督只看单张图，不检查三维自洽	用原生三维路线工具；或用多视角流程约束一致性
拓扑混乱	均匀三角"芝麻面"、孔洞、非流形边	表面提取算法只保证"像"，不保证"结构对"	重拓扑（主角）或减面（远景）
面数膨胀	单件道具 50 万面，场景卡顿	模型优先外观，不以效率为目标	生成时设面数上限；下载后重拓扑/减面
薄壳与透明崩坏	双层壳、破洞、玻璃像塑料	开放曲面是场式方法的死角；透明不符合多视角一致性假设	改写提示词为实体词；加厚；材质演透明
水密失守	切片器报非流形/开放边/自相交	生成管线不以保证水密为目标	选带水密参数的工具；切片验证；逐项修复
光照烘入材质	模型带着"焊死"的阴影与高光，换光照就不对	观测 = 反照率 × 光照，模型分不开两者	重做/修正贴图；夜景场景慎用自带光照的贴图
纹理幻觉与不一致	色偏流动、接缝错位、背面花纹是正面的复制	各视角独立推断，没有"同一物体"的全局约束	环绕截图核对；接缝与颜色单列检查项
尺度 / 轴向 / 原点错误	差 100 倍、躺着进场、摆放诡异	导出与目标软件的约定不同且未校正	导出检查清单第 1–3 条 + 回导入验证

三份病历放在一起看，有一条清晰的共同线索：所有失败都发生在"模型没见识过的地方"或"约定没有对齐的地方"——要么是数据里没有（背面、薄壳、透明），要么是参数里没选对（面数、水密），要么是格式里没说清（尺度、轴向、原点）。这给创作者的动作指南非常朴素：**数据不足处靠多给输入，参数**

不合意处靠显式设置，约定不明处靠检查清单与回导入验证。没有一条需要"祈祷这次抽卡手气好"。

4.9 动手实践：给三个道具做"出厂检验"

这一节把全章知识装进一条工作流：**生成 → 环绕质检 → 修补 → 重拓扑/整理 → 导出检查 → 回导入验证 → 归档。**三个实践对应三个道具，也对应三种典型的选用场景：文字生成（楼台）、图像生成 + 参数实验（扁舟）、以及一个"故意不用 AI"的对照（玉盘）。

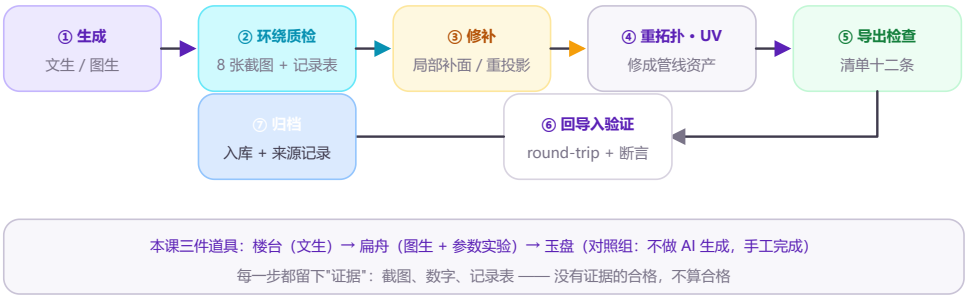


图 4-12　三维资产工作流。从生成到归档的七步，重心明显偏向右侧——"做出来"只要一步，"验出来"要五步。这正是本章想让你带走的工作方式。

🛠 实践一：文生三维 → 多视角质检（50 分钟 • 课堂）

1. **配提示词。**用下面的提示词卡为"临江楼台"写一段生成描述。注意四段式：主体 → 结构细节 → 材质风格 → 用途约束。
2. **生成。**任选一家云端服务（4.7 节表 4-15），用量产设置生成 3 次（同一提示词），保留 3 份结果——这会让你看到"同一提示词的方差"。
3. **环绕截图。**选最像的一个结果，在预览窗口里固定相机高度与距离，每转约 45° 截一张，共 8 张（正、背、左、右、四斜角）。文件名规范：`楼台_v1_视角02_右前45度.png`。
4. **逐项填写缺陷记录表**（表 4-17 模板）：每条缺陷必须配"截图编号 + 严重度 + 处置决定"。
5. **决策。**对每条缺陷给出处置：忽略（不影响用途）/ 局部修补 / 重新生成。写下你判断的依据——这比判断本身更重要。

基础

完成一次完整的"生成 → 8 张截图 → 缺陷记录表（至少 5 条记录）"流程。

提交物：8 张截图 + 缺陷记录表 + 一句话结论（这个模型能不能进第 7 章的夜江场景？为什么？）。

进阶

做"提示词对照"：对比两次生成的记录表，找出"提示词里加了什么词使哪类缺陷变少/变多"（如加入"对称""四个立面一致"等约束词），写 100 字以内结论。

挑战

把最严重的一条背面缺陷做局部修补（补面 + 重投影或在建模软件里手补），修补前后各截同视角一组图，作为"修复证据"附进档案。

✎ 提示词 · 文生三维（楼台初稿）

一个**[唐代风格]**的临江**[两层木结构楼台]**，歇山顶，深出檐，翘角，四面均有木格窗与回廊栏杆；**[对称结构，四个立面风格一致]**；青瓦、深色木质，**[水墨淡彩质感，表面为实体材质]**；作为**[夜景短片背景道具]**使用，比例真实（楼台高约 **[6]** 米），面数控制在 **[中低档]**，模型水密。

面数档位 中低（对照实验里再调高）　水密 开（若工具有该参数）　姿态 默认　参考图 无（纯文字）

图 4-13　主线案例的三件道具（三维资产目录页）。小舟给出题、楼台作主力、玉盘当对照——三件道具对应三种资产策略（AI 生成 / AI 生成 + 参数实验 / 该手作就手作）。它们在同一个浅灰「资产库」背景上并排，是交给下游（第 5、7 章）前应有的样子。 AI 生成

🛠 实践二：图生三维 + 参数对比实验 + 手作对照（70 分钟 • 课堂）

1. **准备输入图。**从第 2 章的扁舟分镜里挑出侧视角度最清晰的一帧作为参考图，并补两帧（正面、俯视）做成"三视图输入包"。如果你的工具只支持单图，那就分两组：单图组 vs 三图组。
2. **参数对比实验（关键环节）。**用同一张参考图，做至少 4 轮生成，每轮只改一个参数：轮 1 默认；轮 2 提高面数上限；轮 3 打开/关闭水密；轮 4 单图 vs 三图输入。每轮同样做 8 视角截图（可只截 4 个代表视角以省时间），把"参数 → 观察结果"填进表 4-18。
3. **结论句。**为每一轮写一句结论，格式统一："把 X 从 A 改成 B，观察到 Y 变化，说明 Z。"
4. **玉盘对照实验（反向教学）。**用"玉盘 / 圆盘 / 白色瓷盘"做一次生成；同时，在建模软件里用"旋转成型"（画一条剖面曲线绕轴旋转）手工做一个圆盘——熟练者约 5 分钟。对比两者：成品质量、可用性、所花时间。把结论写进档案：简单回转体，手工比你想象的快，也比你想象的干净。

进阶

完成 4 轮参数实验 + 玉盘对照实验，产出表 4-18 的完整记录与两条以上的可复用结论（例如"哪些参数值得每单都调一次"）。

挑战

给扁舟做一次"薄壳专项质检"：按 F4-2 的处方，先检查船篷是否双层壳/破洞，再决定是否加厚与修补；写一段 150 字以内的"船篷专项报告"。

🛠 实践三：导出检查与回导入验证（40 分钟 • 课堂）

1. **导出。**把修好的楼台或扁舟导出为两种格式：FBX（引擎方向）与 GLB（评审方向）。导出记录下单位与轴向设置。
2. **回导入验证。**新建一个干净的场景（Blender 或引擎工程皆可），导入 FBX，逐条执行表 4-12 的前 8 条检查；能量化的量化（量尺寸、数面数、数非流形边），留截图。
3. **填写"交接单"。**一页纸写清：资产名、用途、面数、尺寸、单位、格式清单、已知缺陷、缺失项（碰撞/LOD）、验证人、日期。
4. **交叉验收。**与隔壁组交换资产做"陌生环境导入"——让没碰过这个文件的人按交接单复现检查。凡是对方看不懂或复现不了的条目，回去重写。

基础

完成一次完整导出 + 回导入验证，12 条清单全部留下证据。

进阶

在检查中发现并修复至少 2 个问题（如尺度、原点、通道），把"问题—修复—复验"三段写成记录；体会"约 90% 的导入事故可以被清单提前捕获"这句话在你自己手上的成色。

表 4-17　缺陷记录表（模板，复制到创作档案）

编号	发现视角	缺陷类型	严重度	截图编号	处置决定	状态
01	背面	背面糊面 / 结构坍塌	高	楼台_v1_视角05.png	局部补面 + 重投影	已修复
02	右后斜角	多余结构（无观测来源）	中	楼台_v1_视角06.png	剪除并补面	已修复
03	正侧	贴图接缝错位	低	楼台_v1_视角02.png	记录，暂不处理（远景不可见）	搁置
04	全部	面数超标（约 42 万）	高	—（数据）	重拓扑到 8,000 面档	进行中
05	俯视	顶部法线反转（黑面）	中	楼台_v1_视角07.png	翻转法线	已修复

表 4-18　参数对比实验记录（模板）

轮次	改变的唯一参数	观察到的变化	结论句
1	—（默认）	正面好、背面糊；面数约 30 万	默认档适合快速看形体，不适合直接交付
2	面数上限：中 → 高	细节增加但模型更重；缺陷类型不变	提高面数不治幻觉，只加成本
3	水密：关 → 开	孔洞减少；打印验证通过率上升	下游是打印时，水密开关优先级高于面数
4	输入：单图 → 三视图	背面结构明显改善；多余结构变少	"多给视角"是性价比最高的一步

🎯 归档规范：证据留档四件事

每件道具归档时，四样东西一起放进档案：① 生成记录（工具名、提示词、参数、日期、额度消耗）；② 质检证据（8 视角截图 + 缺陷记录表）；③ 导出记录（格式、单位、轴向、回导入验证结果）；④ 来源与许可记录（参考图出处、工具输出的许可类型、署名要求）。少任何一样，事后就无法复现、无法追责、也无法商用。第 4.11 节会告诉你，第四样在"拿别人的照片做图生三维"时会变得多么重要。

4.10 关键概念表：本章词汇的"速查手册"

本章新词密集，把它们的"一句话定义"集中列在这里。建议用法：**先盖住右列自测**——看到术语想定义，想不起来的回查正文对应章节。

表 4-19　本章关键概念速查

概念	一句话定义	你要记住的那一点
三维表征	用某种数据结构描述物体形状的方式	没有万能表征，只有对用途最划算的表征
点云	一堆带坐标（可带颜色、法线）的点	无连接关系；渲染有空洞；扫描的第一步
网格	顶点 + 面 + 顶点属性组成的表面外壳	工程生态最成熟；AI 交付的最终形态通常是它
体素	规则立体方格网，每格记录"有/无"	存储随分辨率三次方增长（n^3）
隐式场 / SDF	记录"每个位置距表面多远、在内还是在外"的函数	表面光滑、可布尔运算；要"看见"必须取样提取
神经辐射场（NeRF）	沿光线查询"位置+方向 → 密度+颜色"的可渲染函数	真实感高；但慢、难编辑、无显式几何
三维高斯泼溅（3DGS）	用一团团半透明椭球表示场景	实时渲染快；有悬浮物、几何不能直接当墙用
重建 / 生成	复现真实物体 / 合成不存在的物体	分界线是信息量；生成本质是"用先验补观测"
SDS 分数蒸馏	用二维扩散模型当评委，把它的"修改意见"蒸馏给三维表示	证明"无三维数据也能造三维"；但慢且易幻觉
多面神（Janus）	每个方向看都像正面，转一圈看到多张脸/多个同款	二维监督不检查三维自洽的必然产物
原生三维生成	直接训练生成三维结构的大模型（稀疏体素 + 结构化潜变量 + 流匹配）	秒级出模；但训练分布的边界就是它的边界
拓扑	网格的连接方式（谁连着谁）	外观看不出，决定能不能绑骨、能不能打印
流形 / 水密	每条边最多两个面（流形）；完全封闭无洞（水密）	打印与物理的前提；生成模型不保证这两点
面数预算	一件资产允许使用的三角形数量上限	是预算不是军备竞赛；AI 直出常超标十几到几百倍
重拓扑 / 减面	重建一张好网格 / 机械删除三角形	主角用重拓扑，远景用减面
UV 展开 / 接缝	把曲面剪开摊平成贴图坐标 / 剪开的位置	接缝藏天然分界；先重拓扑再展 UV
法线贴图 / 烘焙	不增面数而"画"出凹凸的贴图 / 把高模细节转移进贴图	AI 结果当高模、重拓扑出低模、烘焙回细节
PBR / 金属-粗糙度工作流	用物理参数（颜色、法线、粗糙、金属等通道）描述材质	通道打包与 sRGB 约定是导出事故高发区
确定性断言	能用程序判定真假的检查（缩放=(1,1,1)、非流形边=0…）	截图只证明外观；断言才证明资产
回导入验证	把导出文件导回干净环境重跑检查（round-trip）	导入决定出货；这是把关的最后一关

4.11 课程思政：文化遗产数字化复原的边界与真实性

本章的技术有了一种特殊的用途，也因此有了一种特殊的责任。三维生成能"补全"缺失的一切：缺一角飞檐，它给你补上；少一段史实，它替你想象。对普通道具，这叫效率；对文化遗产，这叫危险——因为观众会把你"补"出来的东西，当成"它本来就是这个样子"。这一节讨论的就是这条红线：技术向文化致敬的方式，是无限逼近真实、并诚实地标出"哪一步是推测"。

设想这样一个课堂场景（虚构情境，用于讨论）：小组决定给短片加一段"古城意象"，小月提出——能不能顺便做一套"某某古桥的数字复原"，发布时说"这就是它当年的样子"？老师的回答可以浓缩成一句话：**你可以复原，但不能代替历史说话。**"复原"是一个有严格含义的词，它意味着你的每一个决定都有证据链；把"我以为"包装成"本来就"，就不再是创作，而是对公众历史认知的污染。

◆ 4.11.1 三层边界：复原、推测、演绎

行业与学术界的通行做法，是把"数字重建"按证据强度分成三层，每一层对应不同的自由度和不同的标注义务。请把这张图当成你做文物类项目时的"交通规则"：

图 4-14　数字复原的三层边界。有据复原、合理推测、艺术演绎——三层都可以做，但标注义务逐层加码。最危险的错误不是"推测"，而是把推测说成"复原"。

三层边界对创作的实际约束，可以落到四个具体动作上：**第一，先查后做。**动手生成之前，把"哪些信息有依据、哪些没有"列成清单——这份清单就是你后面所有标注的源头。**第二，分区建模。**有据的部分按测绘数据建，推测的部分单独成组并命名（如 `推测_屋面`）——命名本身就是证据管理。**第三，交付即标注。**在任何形式的发布里（视频、演示、网页）标注"复原 / 推测 / 演绎"的性质与依据；短片的片尾滚动字幕里加一行"本片古建形象基于 XX 材料复原，屋面部分为推测"，也许没人细看，但它是创作者专业与否的勋章。**第四，不做"以假乱真"。**不把 AI 复原用作"实物照片""官方资料"性质的展示；为文物"数字化"的产品若涉及商业用途，还要回到下一小节的授权问题。

◆ 4.11.2 规则在哪里：与三维创作直接相关的几条法定要求

上面讨论的是行业与学术的伦理规范；除此之外，还有一条更硬的底线——法律。下面这张表把与"三维生成与编辑"直接相关的规则集中列出（本章截稿时可查的主要规则，法规会更新，引用时以官方发布为准）：

表 4-20　与三维创作直接相关的规则（摘要，非法律意见）

规则	关键内容	对三维创作意味着什么
《民法典》第 1019 条	未经同意，不得制作、使用、公开他人肖像	用真人照片做图生三维、做人脸重建，先取得同意
《个人信息保护法》第 28–30 条	生物识别信息属于敏感个人信息；处理须取得**单独同意**并告知必要性与影响	人脸、人脸模型都与"生物识别"相关，授权要求高于普通照片
《互联网信息服务深度合成管理规定》（2023 年 1 月 10 日施行）	提供人脸、声音等生物特征编辑能力需提示取得被编辑人单独同意（第 14 条）；特定情形做安全评估（第 15 条）；对合成内容加技术标识、对生成内容作显著标识（第 16、17 条）	你用的"图生三维人脸"工具与你的作品发布，都在其射程内
《生成式人工智能服务管理暂行办法》（2023 年 8 月 15 日施行）	第 12 条：提供者应对图片、视频等生成内容进行标识	发布 AI 生成的三维渲染图/视频，要带生成标识
《人工智能生成合成内容标识办法》	要求显式标识（看得见的提示）与隐式标识（文件元数据）并行	给渲染产物加"AI 生成"角标 + 元数据标识，双保险
最高法《关于依法审理涉人工智能纠纷案件的意见》（2026 年 9 月 7 日，法发〔2026〕10 号）	五部分 24 条；对人格权做三层保护；侵权判断以"可识别性"为标准，而非"完全一致"	不必"长得一模一样"才侵权——能认出是谁，就可能落在保护范围里

把表格翻译成一句给创作者的话：**凡涉及"真人肖像与人脸/声音"的三维资产，授权要求是最严的一档（单独同意）；凡涉及发布，标识义务是最基本的一档（显式 + 隐式）。**注意最高法意见里"可识别性"三个字的分量：它不是问你"这张脸是不是像素级复刻了某人"，而是问"观众能不能认出这是谁"。对三维创作者的直接推论是——把同学的脸生成成角色、把老师的形象做成手办模型，即便"做得很不像"，只要能被认出，就已经踩在了别人的权利边上。课堂练习的安全做法：用虚构角色，或让"模特"本人签署知情同意并全程参与。

◆ 4.11.3 资产许可：数据与输出的"来源链"

文化遗产之外，还有一类更日常的"真实性"问题：你生成的东西，权利链是干净的吗？三维生成站在一串长长的来源链条上——训练数据来自网络图像与三维库，你的产出又建立在模型的许可条款之上。链条上任何一环有瑕疵，都可能在你最意想不到的时候（作品获奖、视频走红、客户验收）找上门。

三条从材料里来的现实提醒，请务必消化：① 大型三维数据集的名义许可常常不等于逐件许可——例如 Objaverse-XL 整体标称 ODC-By v1.0，但**逐对象的许可并不一致**，其中 Polycam 子集明确只限学术非商用。即"数据集允许"不代表"里面每一件都允许"。② 国产大模型的开源/免费不等于无约束——混元 3D 的社区许可有地域排除（不含欧盟、英国、韩国）、商用规模门槛（以百万月活为线）以及"禁止用输出改进其他 AI 模型"等条款；而 Tripo、Meshy 的免费层输出则是 CC BY 4.0——**允许商用但必须署名。**③ 生成模型本身的训练来源正被推上法庭与考据台：材料记录了 2026 年针对 Meta 的诉讼（涉及其三维生成模型），法院在 2026 年 8 月下旬驳回了其中两项主张但未触及实体问题，同时还有针对其他大厂的平行诉讼的指控——结论不是"不能用"，而是"要在用之前读清楚、记下来、能出示"。研究界也已出现专门追查"某个三维模型输出源自哪个生成器"的被动式来源归属基准（首次覆盖 22 个三维生成器）——这说明整个生态正在快速装上"溯源"能力：你交付的资产，将来是要能被追溯的。

由此得到商用前的"三查"动作，简单但有效：**一查输入**——参考图/参考模型是哪里来的、有没有授权（尤其是真人、知名 IP、他人作品）；**二查工具**——所生成资产的许可类型、署名要求、商用限制（把这些条款原文截图存档）；**三查输出**——发布时是否按要求标识、是否附带署名与来源说明。

> 🏛 **课程思政** 与历史相处的四条纪律
>
> 把本节收束成四条可以直接贴在工位上的纪律——它们是本课程对"技术赋能文化"的完整立场：
>
> ① **有据才敢言复原。**复原是一座需要证据链支撑的桥：测绘、文献、考古材料——每一块砖都要说得出来源。缺证据的部分，归入"推测"或"演绎"，绝不混装。
>
> ② **推测要亮明身份。**推测不丢人，隐瞒推测才丢人。以浙江为例，无论你做的是良渚遗址的展示片段，还是江南古镇的数字街景，凡属推演补全的内容，都在作品与说明中显著标注"此为推测性复原及依据"。**观众有权知道自己看见的是"证据"还是"想象"。**
>
> ③ **对人要有敬畏，对物要有出处。**真人形象（含生物特征）需单独同意；文物与非遗素材需遵守管理要求；模型与数据的许可需逐条核对。技术让你"能做"，许可与伦理决定你"可不可以做"。
>
> ④ **生成必标识，传播必诚实。**AI 生成的内容主动标识，不冒充史实、不冒充实物照片、不冒充他人作品。你做的每一帧画面都在参与公众的历史认知——技术越强，越要把"我不知道的部分"留在明处。

四条纪律背后的精神，其实是同一件事：**把技术的力量用在"更接近真实"，而不是"更方便地制造想当然"。**我们这一代人很可能是第一批"能用一句话造出一座楼台"的人；也正因如此，我们怎么对待"不知道的部分"，会成为后来者理解传统的方式。

4.12 本章产出与下一站

收束主线：这一章里，春江小组把三个道具从"念头"推进到了"台账"。

图 4-15　三件道具的交付台账。三条路线、三种结局：AI 生成主力道具（楼台）、AI 生成 + 参数实验（扁舟）、以及"该手工就手工"的对照（玉盘）。这张台账就是你们组这一章的成绩单。

回看这一章的路：我们先把三维拆成六种表征理解了它"长什么样"；再学会分清"重建"与"生成"两条来路；然后看穿了两代生成技术的底牌——SDS 如何用二维老师教三维学生、原生三维如何秒级出模，又各自在哪儿露怯；接着握住四个"预览里看不见"的工程抓手（拓扑、面数、UV、PBR），并用"导出检查清单 + 回导入验证"把资产送过交接线；最后在思政一节里立好了与历史、与他人、与许可相处的坐标。三个道具在台账上各就各位——它们的下一段旅程有两种走法：要么被摆进第 7 章的夜江场景当静物，要么被装上骨骼去演戏。先去做后者：**让它们动起来。**

◆ 把这一章压缩成七句话

1. 六种三维表征（点云/网格/体素/隐式场/NeRF/3DGS）没有最优解，只有对用途最划算的取舍；而本书所有资产的最终形态都必须是网格。
2. "重建"复现实物、可量化验收；"生成"创造新物、只能主观评价——分界线是信息量，信息不足处模型只能凭先验"编"，一切幻觉由此而来。
3. SDS 用二维评委教三维学生：慢、易多面神、监督失配；原生三维用稀疏体素 + 流匹配把生成拉到秒级，但训练分布的边界依然是它的边界。
4. 评价的两把尺子并不同步：人评在奖励渲染格式与贴图，而不是几何正确性——所以质检要靠"环绕截图 + 缺陷记录表"这两件土兵器。
5. 拓扑、面数、UV、PBR 这四件事在预览窗口里全都看不见，却决定资产"能不能用、在不在预算内、皮肤贴不贴得住、换到哪儿都不塌"。
6. 交接的三坑是尺度、轴向、原点；清单十二条能捕获约 90% 的导入事故，而回导入验证（round-trip）是把关的最后一关——导入决定出货。
7. 技术越强，质检越贵；对文化遗产，复原要有据、推测要亮明、演绎不冒充、生成必标识——技术向历史致敬的唯一方式是诚实。

◆ 自测

❓ 自测 1：为什么说"没有一种三维表征是万能的"？请用"要做物理碰撞的游戏场景"与"要照片级真实感的虚拟导览"两个需求，各选一种表征并说明理由。

答案：因为六种表征各有强项与死角：数值精度、视觉保真、内存开销、渲染速度、可编辑性无法同时拉满。游戏场景需要"能撞、能走、能被摆"的显式几何——选**网格**（渲染快、工具链成熟、可绑骨骼、有碰撞体），这也是为什么本书项目的资产最终都要转成网格；虚拟导览追求"看起来真"且允许渲染方式特殊——可选 **NeRF（离线）或 3DGS（实时）**，代价是没有可用的碰撞几何、编辑与导出受限。判断口径永远是三个问题：要不要交互？要真还是要准？后面还改不改？

❓ 自测 2：小舟说"我用文字生成了楼台，正面很好看，直接拖进场景吧"。请列出他在这一句话里至少违反了本章的哪三条规定动作。

答案：① **没做多视角质检**——"正面好看"恰恰是最不可信的验收视角，背面是背面幻觉的高发区，必须用环绕截图逐张核对；② **没做数据检查**——面数、拓扑、水密、UV 这些"预览里看不见"的项目一条都还没查，场景卡顿或导入报错就在路上；③ **没做交接检查**——尺度、轴向、原点、单位（以及清单十二条与回导入验证）未过，模型"大 100 倍、躺着进场"正是这类疏忽的典型后果。正确顺序：先质检与修补，再重拓扑/整理，再导出检查，最后回导入验证。

❓ 自测 3："重建"与"生成"的分界线是什么？以下三个任务分别属于哪条路线：① 用手机围着石狮拍 60 张照片做成模型；② 用一张水墨分镜图生成扁舟；③ 为不存在于史册的"唐代临江楼台"产出三维初稿。

答案：分界线是**信息量**：观测足以支撑几何 → 重建；观测不足、需要模型用先验补 → 生成（或混合）。① 属于重建（多视角观测充分，有标准答案可量化验收）；② 属于生成（单图信息不足，背面等处靠模型推断，务必要求多角度质检）；③ 属于生成（目标物不存在，没有标准答案，只能凭"合理性"评价）。注意②的边界：如果拍摄的是真实船模的 60 张照片，它就变成了重建。

❓ 自测 4：SDS 路线的"多面神"问题是怎么产生的？为什么有报道说"某改进把质量指标 3D-FID 从 75.96 降到 58.75"，却不能说"多面神问题解决了"？

答案：多面神（Janus）的根源是**二维监督只对单张图负责**：SDS 让二维扩散模型对每一个渲染视角分别打分，从不检查"这个视角与那个视角是否指向同一个三维物体"，于是模型可以学到"每个方向都像正面"。3D-FID 之类的指标衡量的是单视角观感质量，与"三维自洽"是两把尺子；材料中 ProlificDreamer 的 3D-FID 明显优于 Magic3D，但多面神比例仍达约 40%——质量指标的改善来自更稳的蒸馏信号，并不自动产出自洽性。要真正缓解，要么改换原生三维路线，要么在流程上强制多视角一致与人工验收。

❓ 自测 5：小舟的水杯导出后 50 多万面，进切片软件被报"非流形边"，接着又暴露"开放边"，修复后壁被删掉。请按"成因—检查—修复"的顺序给出完整处置方案。

答案：成因：生成模型的几何来自表面提取，天然可能带孔洞、非流形结构与自相交，同时面数超预算数十倍——它被训练成"优先外观、不顾效率与水密"。**检查：**按序跑切片的报错源（非流形 → 开放边 → 法线翻转 → 壁厚）、给面数与自相交比例留下数字证据，同时用环绕截图确认外观损失可接受。**修复：**① 面数：重拓扑（要编辑）或减面（只围观）；② 拓扑：修非流形与开放边、统一法线；③ 壁厚：加厚到打印机极限以上（FDM 约 1–1.2 毫米，SLA 约 0.6–0.8 毫米），或改写提示词/参数从生成端规避（必要时选"水密"开关重生成）；④ 复验：重新切片通过 + 回导入验证通过才算出厂。要点是"按序修、留证据、复验"，而不是反复导出撞运气。

❓ 自测 6：小组想用 AI"复原"一座地方古桥并作为短片发布。请从"真实性"与"合规"两方面，列出发布前必须完成的动作清单。

答案示例：真实性方面：① 先查后做——测绘/考古/文献依据列出清单，区分"有据、推测、演绎"三层；② 分区建模并命名（如"推测_屋面"），让证据可管理；③ 在成片与说明中显著标注各部分的性质与依据；④ 不把复原结果用于冒充实物照片或官方资料的场合。**合规方面：**⑤ 若素材涉及真人（照片、人脸、声音）——取得单独同意，管住生物特征信息的使用范围；⑥ 逐条核对所用工具与数据的许可（地域、商用门槛、署名要求），原文存档；⑦ 发布时按《标识办法》做显式（可见标识）+ 隐式（元数据）标识；⑧ 如涉文物主管单位的资料或场景，遵守其管理要求并保留授权记录。最后一条建议：把以上动作写成一张《发布前核对单》，交付时与作品一起归档——这既是自我保护，也是专业主义的标志。

◆ 拓展阅读

- **下一章：**第 5 章 角色动画生成与编辑——骨架怎么装进网格、动画怎么被生成与编辑。本章表 4-13 的"绑骨前检查"就是它的入场券。
- **场景应用：**第 7 章 游戏场景生成——道具如何被摆进场景、光线如何统一，你的"楼台/扁舟/玉盘"将在那里完成布景。
- **工具手册：**Blender 官方手册的"UV 展开""重拓扑""PBR 材质"章节——本章实践二、三的操作说明书。
- **标准与规范：**Khronos glTF 规范（材质与通道打包的权威出处，对照表 4-10 阅读效果最佳）。
- **事实来源：**本地研究报告《三维模型生成：路线、瓶颈与评测的事实地图》及各工具官方页面——本章所有参数与许可条款的核对起点；记住"三源核验"：官方文档 > 官方页面 > 二手转述。

第 5 章

角色动画生成与编辑

让诗人举起酒杯：骨骼、蒙皮与关键帧，把静止的模型变成会演戏的角色。

学习目标

1. **知识：**理解骨骼（armature）、蒙皮（skinning）、权重、关键帧、插值曲线、IK/FK 与动作重定向（motion retargeting）的准确含义；认识动画十二原则中与质量检验相关的几条；理解不同骨架之间"翻译"动作的基本原理。
2. **能力：**能对第 4 章导出的三维模型做"绑定前体检"；用网页工具完成自动绑定（auto-rigging）并判断蒙皮权重的好坏；用动作库或文生动作（text-to-motion）得到一段动作并套用到自己的角色上；能用关键帧做人工修正；能诊断并修复常见的网格与材质缺陷。
3. **素质：**建立"AI 出草稿、人做判断"的创作观，理解人工修正的审美价值与工匠精神；尊重动捕数据与动作素材的授权边界，恪守 AI 生成内容的标识要求与合规底线。

对应第 6 周 · 3 学时　讲授 1 学时 · 实践 2 学时　主线推进：诗人「举杯邀月」

> 第 4 章的末尾，小舟从三维工具里导出了诗人角色的模型：一个安静站在 T-pose 里的人，衣袂整齐、五官俱全、贴图漂亮——但他**不会动**。小月把模型转了一圈，问出了本章要回答的问题："怎么让他举起酒杯，抬头看月亮？"

在三维世界里，"让角色动起来"有两层含义。一层是**技术**上的：网格要能被骨骼带动、动作数据要能被正确套用、导出后引擎要认得。另一层是**表演**上的：动得对不难，动得好看很难——同样是举杯，有人举得像举哑铃，有人举得能让观众相信他真在邀月。这一章我们把两层都走一遍：先讲透骨骼、蒙皮、权重、关键帧、插值曲线、IK/FK 与动作重定向的原理，再看 AI 工具把这条流水线推进到了哪一步，最后回到小月的问题，用"自动绑定 + 动作获取 + 人工修正"三条腿，做出 4 秒钟的「举杯邀月」。请先记住本章反复使用的分工原则：动画的"及格"由技术决定，"好看"由人的判断决定。

5.1 从静止到会动：动画到底在动什么

先给动画下一个能落地的定义：**动画 = 姿势 × 时间。**所谓"姿势"（pose），是某一瞬间角色**所有关节的角度集合**；所谓"动画"，就是这套角度随时间连续变化的序列。屏幕上你看到的"举手"，其实是"肩关节的旋转角度从 0° 变到 90°"这样一串数字在时间里跑了一遍。

这个定义听起来朴素，却直接决定了后面所有工具的分工：既然姿势是"一堆数字"，那么**谁来产生这些数字**就成了关键问题——动画师一个一个摆（手 K），动作捕捉设备录（动捕），AI 根据一句话生成（文生动作），或者把别的角色的动作翻译过来（动作重定向）。本章后面讲的每一种 AI 工具，本质上都在回答这同一个问题。

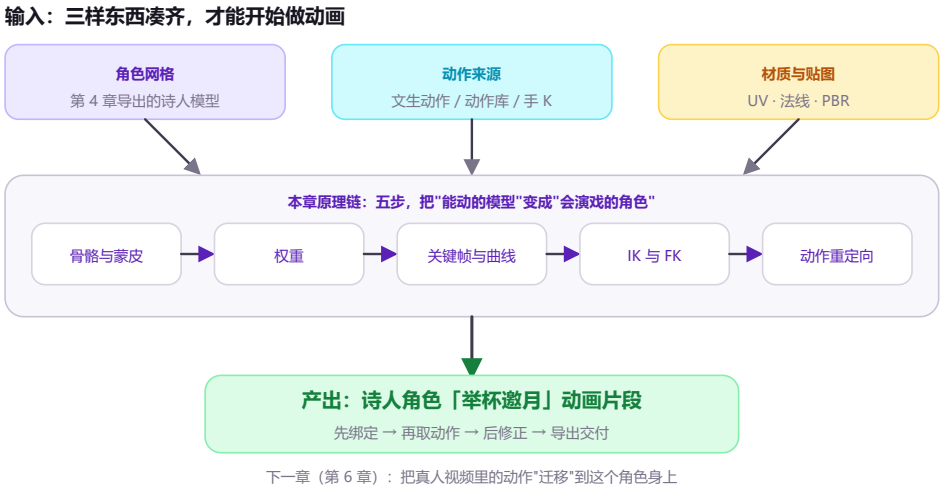


图 5-1　第 5 章地图。本章处理的输入是"一个不会动的角色 + 一段动作数据 + 一套材质"，输出是"一段能演出的动画"。三个输入缺一个，最后都会在某一集卡住——这正是本章既要讲动画原理、又要讲材质修复的原因。

◆ 帧、帧率与"关键帧"

动画的最小时间单位是**帧**（frame）：一张静止画面。一秒里播多少帧，叫**帧率**（frame rate，单位 fps，每秒帧数）。电影与院线动画的经典帧率是 24 fps，短视频与多数 AI 视频工具是 30 fps，游戏为了操作手感常用 60 fps。帧率对制作的含义很直接：同样 4 秒钟的动作，24 fps 只需要 96 帧画面，60 fps 就要 240 帧——帧率越高，动作越顺滑，但数据量与渲染成本也随之上升。

但动画师并不是"一帧一帧地画 96 张"。无论是二维还是三维，行业都用同一个办法：**只定义少数几个"关键姿势"（关键帧），中间的过渡交给计算机补。**这个"补"的正式名称叫**插值**（interpolation）。这里有一个特别值得讲的对照：**在传统二维手绘动画里，中间帧由初级动画师一张张手绘；在三维动画里，中间帧由计算机根据样条曲线自动算出来。**计算机算得比人快无数倍，但它只会"算"，不会"表演"——所以三维动画师的工作重心，从"画中间帧"变成了**摆关键姿势**和**调曲线。**这两个词，就是本章后半段"人工修正"的全部内容。

表 5-1　帧率与实践含义：同一个 4 秒动作

帧率	每秒帧数	4 秒动作总帧数	常见用途	对制作的影响
24 fps	24	96	电影、院线动画、绝大多数"有质感"的短片	经典"电影感"；略低的帧率反而让动作更有"重量"
30 fps	30	120	短视频、网页动画、多数 AI 视频与文生动作模型	本书与课程作业的默认选择；与视频素材对齐方便
60 fps	60	240	游戏实时动画、需要"跟手"的交互场景	数据量与算力翻倍；对短片创作通常没有必要
非整数帧率（如 23.976）	≈24	≈96	历史遗留的影视工程标准	知道有这回事即可；混用整数与小数帧率会造成音画不同步

> **📌 一个容易混淆的说法："12 帧动画"**
>
> 你可能听过"日本动画是 12 帧"这类说法。它的意思不是"帧率是 12 fps"，而是**在 24 fps 的时间轴上，每张画面停留两帧**——即"一拍二"（on twos）。这样一秒里实际只有 12 张不同的画，既省工，又形成了独特的节奏感。三维动画里的"减帧渲染"（先按低帧率摆姿势、再补帧）与它是同一个思路：观众感知的不是"每秒多少张"，而是"动作的快慢与轻重"。

◆ 姿势：角色的"数字状态"

既然动画是"姿势随时间变化"，我们得先知道一个"姿势"在计算机里长什么样。答案很朴素：**每个关节的旋转角度 + 根节点的位移。**以一个人形角色为例，根节点负责"这个人站在哪儿、朝哪边"，其余关节各自记录"相对父关节转了多少度"。

这个数量级你可以心算一下：如果一个角色有 24 个关节，每个关节用 3 个数字表示旋转（对应上下、左右、扭转三个方向），那么"这个人此刻的姿势"就是 72 个数字；再加上根节点的位移，差不多 75 个数字就能完整描述。第 6 章你会遇到的人体参数化模型 SMPL，正是用"24 个关节 × 3 个旋转自由度 = 72 个姿态参数"加上 10 个控制体型胖瘦的形状系数来描述一个人——这不是巧合，而是"用最少的数字描述一具身体"这个工程需求共同的答案。理解了这一点，你就能理解后面所有 AI 动作工具的输出格式：它们生成的不是"视频"，而是**一串这样的数字序列。**

◆ 两条制作路线：手 K 与数据驱动

得到这串数字序列的方法，行业里主要有两条路，理解它们的差别，你就理解了本章所有 AI 工具的定位：

- **手 K（手工关键帧）：**动画师在软件里摆姿势、调曲线，一帧一帧地"表演"出来。它最自由——想演什么就演什么，包括现实中不存在的动作；也最费功夫，质量完全取决于动画师的表演修养。
- **数据驱动（动捕 / 动作库 / 生成模型）：**用动捕设备或算法得到真实动作数据，再套用到角色上。它快、真实，但受限于"数据里有什么"——而且数据与角色之间还需要一次"翻译"（就是本章 5.7 节要讲的动作重定向）。

AI 工具落在这条谱系的"数据驱动"一侧：文生动作（text-to-motion）是"用文字描述去生成动作数据"，动作库是"用现成的动作数据去匹配"，自动绑定则是"把角色准备好以接收数据"。它们共同把门槛从"会做动画"降到"会提需求 + 会验收"，而**验收所需的眼光，恰恰要靠本章前半段的原理来建立。**

5.2 骨骼：角色的"骨架"与运动链

要让网格动起来，第一件事是给它一副"骨架"。在三维软件里，这副骨架是一个独立于网格的对象：**骨骼 / 骨架**（armature；更通用的说法是 skeleton，即"骨架"）。需要说明的是，"armature"是 Blender 对骨架对象的专用称呼，"skeleton"在跨软件语境下更通用——本书两个词都会用到，指的都是同一件东西。

骨架不是一堆散落的点，而是**一棵树**：每个关节都挂在另一个关节下面，形成父子关系（parent-child）。父关节一动，挂在它下面的所有子关节整体跟着动；反过来，子关节动，父关节纹丝不动。这就是"**运动链**"（kinematic chain）——链条上的运动按层级**单向传递**。生活里最好的类比是手臂："转肩膀，整条胳膊一起走；转手腕，只有手掌动。"

图 5-2　骨骼层级与运动链。左图是一条手臂链：转肩带动整条臂，转肘只带动前臂与手。右图是完整的角色骨架——记住"上管全局、下管局部"这一条，后面理解权重、IK、重定向都会顺畅很多。

◆ 关节的自由度与限位

不是所有关节都"想怎么转就怎么转"。这里有两个概念必须分清：

- **自由度**（degree of freedom，DOF）：这个关节允许几个方向上的旋转。髋、肩这类**球窝关节**有三个自由度（上下摆、前后摆、扭转）；肘、膝这类**铰链关节**通常只有一个（只能屈伸）。自由度越多，动作越灵活，也越难自动计算——这也是为什么"手指"一直是动画里最贵的部分。
- **关节限位**（joint limits）：给每个自由度规定"能转到多少度"。膝盖不能向后弯、手肘不能反折、脖子不能转 360°——这些不是软件的限制，而是**身体的物理事实。**

为什么必须专门讲这两条？因为AI 生成动作时最常见的"穿模"与"肢体反折"，根源就是模型不懂人体力学——它只学会了"统计上像不像一个动作"，没有内置"膝盖不能反着弯"这条约束。理解了"限位"的存在，你就知道处方该往哪个方向开：要么在生成端把约束写进运动条件，要么在修正端手动把越界的关节掰回来。

表 5-2　常见关节的自由度与限位（经验值）

关节	类型	自由度	典型限位方向	动画中的注意点
髋 / 肩	球窝关节	3	各方向大范围，但都有"极限姿势"	最容易出现"脱臼感"（肩膀移位）与穿模（手臂插进躯干）
肘 / 膝	铰链关节	1	单向屈伸，不可反折	反向弯曲是 AI 动作最典型的"解剖错误"，一眼就能看出来
腕 / 踝	复合关节	2	屈伸 + 少量侧向	脚踝的角度决定"脚掌是否贴地"，直接影响滑步判断
脊柱链	多节链	每节 3 个，整体受限	累计弯曲有生理上限	躯干僵硬是"橡皮人"的另一种表现；柔顺的脊柱是"有生命"的关键
手指	多节铰链	每指 3 个左右	屈伸为主	数据最稀缺、最贵；动捕数据常常"手指信息缺失"（详见 5.8 节）

◆ T-pose 还是 A-pose：一个被低估的工程决定

角色绑定前摆什么姿势，看起来是小事，实际上是整条工具链的**标准接口。**业界有一条心照不宣的约定：用于绑定与动画的资产，最好以 T-pose 或 A-pose 交付——Rodin 这类生成工具会强制输出 T-pose/A-pose，Tripo Studio 的自动绑定以 T-pose 为标准输入，Meshy 的自动绑定同样把 T-pose/A-pose

当作标准输入。为什么它们如此一致？因为手臂打开时，腋下区域没有自遮挡，关节检测算法能"看清"肩肘的结构，权重计算也不容易被身体与手臂的贴面几何干扰。

两种姿势之间也有细微差别：T-pose（双臂水平）结构最清晰，是许多检测与绑定管线的默认；A-pose（手臂斜向下约 45 度）更接近自然站立，绑定出的肩部权重做日常动作时往往更顺。你不需要为选择纠结，但要知道两条纪律：**其一，绑定前把角色摆到标准姿势，不要"就这么绑"**——姿势不标准，检测出的关节位置会有系统偏差；**其二，绑定之后不要再大改绑定姿态**——绑定姿态是权重计算的参照系，改了它，所有权重相当于被重新解释，形变会集体失灵。

> 🧪 **原理注：为什么"骨架"能带动"皮肤"**
>
> 到这里你可能会问一个很实在的问题：骨架只是一堆线和点，它怎么让**网格**跟着动？答案是：骨架并不"带动"网格，而是**改变网格顶点的位置**。流程是——你先告诉软件"这个顶点归哪几根骨骼管、各占几分"，之后每根骨骼运动时，软件就按着这个约定，把顶点"搬"到新的位置。这套约定叫**蒙皮**（skinning），那份"归谁管、占几分"的清单叫**权重**（weights）。它们是下一节的全部内容。

5.3 蒙皮与权重：网格"听谁的"

有了骨架，下一步是把骨架和网格绑在一起——这就是**蒙皮**（skinning）：把网格的**顶点**绑定到骨骼上，使网格随骨骼运动而形变。绑定发生的那一刻，角色摆出的姿势叫**绑定姿态**（bind pose），工程上通常要求是 T-pose 或 A-pose——手臂张开、没有遮挡，方便软件判断"哪块皮肤归哪根骨头管"。

而**权重**（weights）定义的是这件事的细节：每个顶点受哪些骨骼影响、各占多少比例。为什么需要"比例"而不是"直接指定归属"？因为身体是连续的：肘部那个位置的皮肤，既不属于上臂，也不完全属于前臂——它同时受两者影响，才能在弯曲时平滑地"过渡"。如果每个顶点都只认一根骨骼，手臂弯曲时关节处就会出现折纸一样的尖角。

图 5-3　权重与过渡带。过渡带太窄，关节会"折纸"；太宽，角色会"橡皮"。所谓"调整权重"，调的就是这个过渡带的形状与范围——这也是自动绑定之后最需要人工检查的地方。

权重怎么算出来的？行业里有两条路：**自动计算**（软件按"顶点到哪根骨骼最近、几何上是否连续"等规则批量分配，常用热扩散一类的算法）和**手动绘制**（用一个画笔在模型上"涂"出影响范围，涂红的地方影响大、涂蓝的地方影响小）。自动绑定工具（下一节会集中介绍）省掉的正是第一步——但自动计算出来的权重几乎不可能一次就对，尤其是腋下、裆部、手指、衣摆这些"几何上互相靠近、功能上却互不相干"的位置。行业里有个专门的说法叫权重**"渗漏"**（weight bleed）：不该管的骨骼把手伸到了别人的地盘上。

◆ 常见误区：权重不是"越平滑越好"

刚学权重的人容易走两个极端。第一个极端是"越平滑越好"：把所有过渡带都涂得很宽，结果角色一动就像橡皮——这正是图 5-3 右侧最下面那条的情形。第二个极端是"越干脆越好"：每个顶点只认一根骨骼，结果关节处出现尖锐的折角，像折纸人。

正确的目标是**该软的地方软、该硬的地方硬**。人体本身就是这样设计的：手肘、膝盖表面覆盖着柔软的皮肤和肌肉，过渡带应该宽一些；而**骨骼突出、皮肤很薄的地方（如膝盖骨正面、手背指节）**，过渡带就该窄。所以专业的权重检查从来不是"看颜色均不均匀"，而是**把关节慢慢弯到极限位置，观察哪里塌陷、哪里尖角、哪里被别处拽动**——这也是本章实践环节的第一个验收动作。

表 5-3　权重质量检查清单（对自动绑定结果的验收，经验清单）

#	检查项	怎么查	不合格的表现
1	关节极限位置是否塌陷	把肘、膝、肩各自弯到最大角度看一眼	关节处出现明显凹陷或尖角
2	是否"牵一发动全身"	抬手、抬腿时观察躯干与对侧肢体	身体被拽歪、另一只手臂跟着动 → 权重渗漏
3	每顶点影响骨骼数是否过多	查看权重的统计信息（多数工具可显示每顶点最大影响数）	一个顶点被七八根骨骼同时影响，形变变得不可预测
4	权重是否归一化	检查权重和是否为 1	形变整体缩小或"缩水"，动作幅度对不上
5	左右是否对称	做一次对称动作（如双臂同时抬起）	左右表现不一致 → 模型本身不对称或权重计算有偏
6	衣物与头发等附属部件	单独检查衣摆、飘带、长发的权重	衣摆被"焊"在腿上，或飘带完全不受控地乱飞
7	面部与手指	做张嘴、握拳等小动作	五官错位、手指粘连——面部与手指通常需要专门的一套绑定

∫ 深入一格 · 蒙皮为什么是"加权投票" 选学 · 数学

最常见的蒙皮算法叫**线性混合蒙皮**（linear blend skinning，LBS）。它的表达式简单得出奇：

$$v' = \Sigma_i\, w_i \cdot M_i \cdot v \qquad (\Sigma_i\, w_i = 1)$$

v 是顶点的原始位置（绑定姿态下）；M_i 是第 i 根骨骼从"绑定姿态"到"当前姿态"的变换矩阵；w_i 是该顶点分给这根骨骼的权重；求和遍历所有影响它的骨骼。

把它翻译成人话：每一根影响该顶点的骨骼，都会提出一个"这个顶点现在应该去哪儿"的方案（$M_i v$），最后按权重做一次加权平均。这和本书第 1 章讲的"注意力 = 加权平均"是同一个数学动作——只是这里加权的是**空间位置**，那里加权的是**语义信息**。

这个式子还顺手解释了两件事。第一，**为什么权重必须归一化**：如果所有 w_i 加起来等于 0.8，相当于把顶点"缩小"了 20%，模型会整体缩水。第二，**为什么权重渗漏会毁掉动作**：一个本该只听手臂指挥的顶点，如果被分到了 0.3 的躯干权重，那么抬手时它会往"躯干给的方案"靠 30%，于是整块布料被拉扯——你在图 5-3 右侧最下面看到的那条"橡皮手臂"，就是这么算出来的。

◆ 面部与手指：两个"绑定特例"

在进入下一节之前，先把两个"特例"交代清楚，免得你在实践中反复撞墙。**面部**是表情肌肉的战场：工程上通常用几十条**混合形状**（blendshape，把"微笑""皱眉"等形状作为变量做加权混合）配合少量骨骼来实现，几乎完全独立于身体绑定；**手指**是关节密度的战场：一只手五根手指、每根三节，加上手腕就是十几个关节，逐指的检测与权重是自动绑定里"最高配"的选项之一。多数通用自动绑定默认只保证手掌整体的姿态，逐指检测要看工具是否提供。

本章任务的策略很清楚：「举杯」这个动作不需要精细的手指与面部表演——整手握住酒杯、表情保持安静即可，把有限的时间预算投给"被观众看到"的部位（肩、肘、腰、衣摆），而不是投给看不见的手指关节。知道能力的边界在哪里，是工程判断的一部分：一个总是"在拇指上雕花、却让肩部橡皮化"的创作者，就是没有做这道取舍的人。

失败诊所 F5-1 抬手，整件衣服跟着走

症状：小舟用网页工具给诗人角色做完自动绑定，随手让角色抬了一下右臂。结果不只右臂抬了起来——右半边衣襟被拽得向前凸出，胸口跟着变形，连左肩都轻微晃了一下。小月看了一眼："像橡皮捏的。""像橡皮人"这个词，从此成了小组内部的固定代号。

成因：典型的**权重渗漏**。自动绑定算法在计算权重时，主要依据"顶点与骨骼的几何距离"和"表面连续性"。诗人的衣袖与躯干在几何上贴得很近，算法没有"生活常识"去判断"袖子该跟手臂走、而胸口不该跟袖子走"，就把一部分躯干权重错误地分给了手臂骨骼。另一个加剧因素更隐蔽：第 4 章生成模型留下的**内部相交面与分离部件**——那些看不见的内部面让算法误以为"袖子"和"胸口"是同一片连续几何，于是放心地让它们共享权重。

处方：按"先修根、后修叶"的顺序做四件事。① **回网格**：先在三维工具里删掉内部面、合并分离部件、确认法线朝外——网格本身不干净，权重永远算不对。② **限缩影响数**：把"每个顶点最多受几根骨骼影响"这个参数调小（工程上通常留 2–4 根），从源头上砍掉远处骨骼的"长臂管辖"。③ **手工补刷**：把渗漏区域用权重画笔重新涂一遍——衣襟刷给"手臂 + 胸腔"，胸口刷回"躯干"。④ **做一次对称性复查**：把两个手臂分别抬起，确认左右表现一致。

5.4 关键帧与插值曲线：动画的"标点符号"

骨架搭好、权重调顺，角色就"能动"了。接下来是真正的动画问题：**怎么让它动得好看？**这一步的战场，从"骨骼与顶点"转移到了"姿势与时间"。

前面说过，动画师不逐帧画，而是摆**关键帧**（keyframe）——也就是**关键姿势**——中间交给计算机插值。那么问题就变成了：中间"怎么走"，由谁决定？答案就在**插值曲线**（interpolation curve）里。这是本章最重要的一个可调项，因为**它决定动作的"重量感"与"生命感"，而不是动作的"正确与否"。**

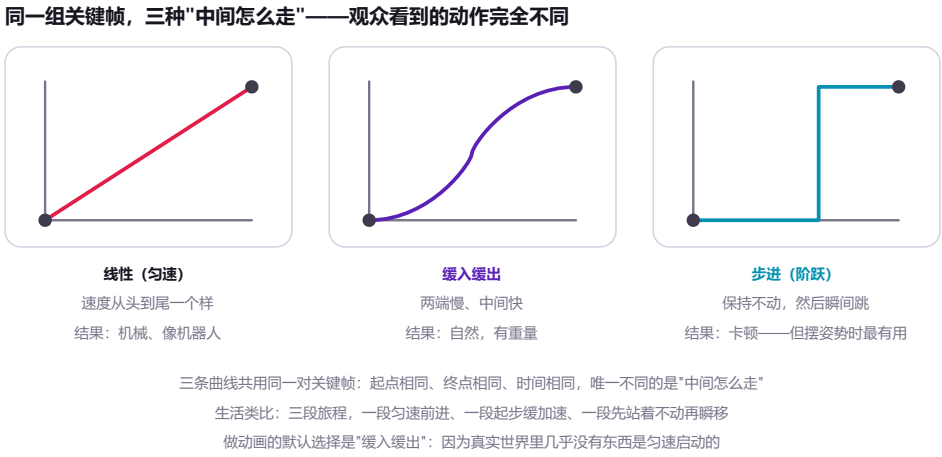


图 5-4　插值曲线决定"动感"。关键帧只规定"经过哪些姿势"，插值曲线规定"怎么经过"。行业经验是：**匀速运动会显得机械**，曲线视图（曲线编辑器）因此成为动画师最核心的工具。

曲线为什么能决定"重量感"？用一个生活中的例子就能讲明白：电梯启动时会先"缓一下"再加速，快到楼层时又会"缓一下"再停稳；而如果没有这个缓冲、匀速直上直下，坐在里面的人会觉得像被机械臂拎着走。**物理世界里，物体从静止到运动、从运动到静止，都需要克服惯性，所以速度变化必然有一个"渐入渐出"的过程**——动画师把这个过程复刻到曲线上，观众的大脑就会自动解读出"重量"。

表 5-4　插值与切线类型速查（工作语言）

类型	它规定什么	典型用途	常见误用
线性 linear	两个关键帧之间匀速过渡	机械运动、传送带、匀速旋转（如车轮、钟表指针）	用在人物动作上——人几乎不做匀速运动，一用就"假"
缓入 ease in	从一个姿势"慢慢启动"，加速进入下一个	物件开始移动、人物起步、重物被推动	用在"急停"上——该减速的地方反而加速了
缓出 ease out	快速接近目标，再"慢慢停下"	物件到位、人物收势、抛出的东西落定	用在"猛拉"上——该爆发的地方显得软绵绵
缓入缓出 ease in out	两端慢、中间快，是最像真人的默认款	绝大多数肢体动作、相机运动	所有段落都用它 → 节奏平淡，没有快慢对比
步进 step	保持当前值不变，到下一关键帧瞬间跳变（阶跃/常数）	Blocking 阶段摆姿势、定格动画、机械的"咔哒"感	忘了在 Splining 阶段换成曲线 → 成片一闪一闪
自动 / 钳制 / 自由 / 平滑 auto / clamped / free / smooth	控制"切线手柄"——即曲线在关键帧处的陡峭程度与方向	精细调节某个关键时刻的加减速	把曲线调出"过冲与回弹"却不知道自己在做什么 → 关节诡异地抖一下

> **🎯 一句话记住曲线的作用**
>
> 关键帧管"到过哪些姿势"，曲线管"怎么过去"。**姿势错了，是表演问题；曲线错了，是节奏问题**——前者要回去重摆关键帧，后者只需要在曲线视图里拖两下手柄。分清这两类问题，你的返工效率会立刻上一个台阶：先判断"错在姿势还是错在节奏"，再决定改哪里。

◆ 网页工具里的"曲线"在哪里

你可能会问：我用网页工具做动画，界面上根本没有曲线编辑器，这些知识还有用吗？有用，而且恰恰相反——**理解曲线的原理，正是为了在"看不到曲线"的工具里也能判断问题。**网页工具通常把插值简化成"关键帧列表 + 时间条 + 一两档缓动开关"，你无法直接拖手柄，但你知道三件事：其一，任何一个关键帧的时间位置都在参与"节奏"的表达——把它前移或后移一两帧，观感就变了；其二，"缓动"开关的有无，决定了这段动作是否匀速——而匀速就是机械感的来源；其三，当你发现工具不给曲线控制时，最小可行的替代方案是**加中间关键帧**：想要"慢起"，就在起点后不远处补一个离起点很近的姿势；想要"快冲"，就让两个关键帧挤在一起。曲线调的是"数值怎么走"，加关键帧调的是"数值在什么时候取什么值"——两者殊途同归，只是前者更省事、后者更笨但哪里都能用。

◆ 三段流程：Blocking → Splining → Polishing

有了关键帧和曲线，还需要一套"先后顺序"。行业里成熟的三段划分是：**Blocking**（关键姿势布局）→ **Splining**（在曲线编辑器里把步进式关键帧切换为曲线插值并清理曲线）→ **Polishing**（打磨、添加重叠与次级动作）。这套顺序在 Blender、Maya、3ds Max 的动画教学里高度一致，也是本章实践环节要照做的流程。

图 5-5　Blocking → Splining → Polishing。这套顺序的价值不在于"流程正确"，而在于它把"表演判断"和"技术打磨"分成了两道关：靠前的一道考验你的眼光，靠后的一道考验你的耐心。

表 5-5　三段流程的目标与检查点

阶段	这一阶段要回答的问题	用什么插值	通过标准（自查）
Blocking 布局	这个动作"是什么"？观众能不能一眼看懂？	步进（step）	每个关键姿势截图出来都成立；去掉时间只看静帧，仍能读出情绪
Splining 顺曲线	这个动作"快慢对不对"？哪里该慢、哪里该快？	曲线（缓入缓出为主）	没有匀速段落；关键处（发力、停顿）有明确的快慢对比
Polishing 打磨	这个动作"有没有生命"？有哪些细节在出卖它？	曲线 + 附加动画	衣摆与头发有延迟感；无穿模、无滑步、无高频抖动
循环与交接	这段动画能不能接上别的片段？	出入范围（循环/往返）	首尾帧姿态可衔接（循环动作）；导出格式与命名符合下游要求

> **🧪 原理注：摄影表与循环类型**
>
> 除了曲线，动画软件里还有一个常被忽视的工具叫**摄影表**（Dope Sheet）：它不显示曲线的形状，只显示"每个关键帧落在第几帧"。它的用处是**整体调时间**——比如把一段 12 帧的动作整体拉长到 16 帧，动作的姿势完全不变，但节奏慢了三分之一。此外，循环类动作（走路、呼吸、待机）还要设置出入范围类型：**循环**（Cycle）、**往返**（Ping-Pong）等。这些概念在 3ds Max 一类软件的动画教学里是标准内容，你在网页工具里遇到"循环播放""镜像重复"这类选项时，对应的正是它们。

5.5 IK 与 FK：两种"让手动起来"的方式

摆姿势时，动画师面对一个根本性的选择：**是转关节，还是拖手？**前者叫 **FK**（正向运动学，forward kinematics），后者叫 **IK**（反向运动学，inverse kinematics）。名字很唬人，但用一句话就能说清：FK 是"我转关节，手自己跟过去"；IK 是"我拖手，关节自己转过来"。

FK 的思路和你在 5.2 节学的运动链完全一致：转动肩膀 → 上臂动 → 带动前臂 → 带动手，运动从链条根部逐级向外传递。动画师需要自己在心里估算"肩膀转 30 度、肘部再转 45 度之后，手会落在哪里"。IK 反过来：先把"手"这样的链条末端（术语叫**末端效应器**，effector）拖到目标位置，再让软件**反解**——"要让手放到门上，肩膀、手肘、手腕各需要多少角度？"

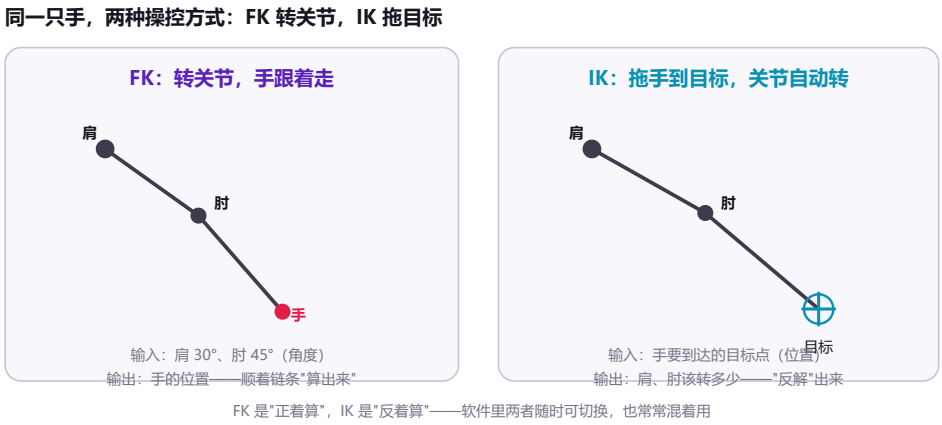


图 5-6　FK 与 IK。FK 从链条根部出发、由角度推位置；IK 从链条末端出发、由位置反推角度。它们不是"新旧"或"高低"的关系，而是两种操控手感，各自适合不同的任务。

那什么时候用哪种？判断标准其实很朴素：**你关心的是"关节怎么转"，还是"手（脚）落在哪"？**挥手、走路时摆手这类动作，动作的"味道"恰恰在关节的角度变化上——用 FK 逐关节调节奏更直接。而"手按在琴键上""脚踩在台阶上""左手扶住栏杆"这类要求末端精确对位的动作，用 IK 把末端"钉"住再动身体，会省掉大量试错。传统动画里有一条经典手法叫"固定末端"：先把脚"钉"在地上，再让髋部与上身表演——这在三维里通常就是用 IK 或专门的"脚锁定"功能实现的。

表 5-6　IK 与 FK 对照：怎么选、怎么用

维度	FK 正向运动学	IK 反向运动学
操控对象	每个关节的旋转	链条末端的（位置与朝向）
运动传递	上动下随：转肩带动整条手臂	下动上随：拖手，肩肘跟着"想办法"
适合的任务	挥臂、摆手、走路的腿部摆动、一切看重"角度节奏"的动作	手扶、脚踩、按压、抓握——一切要求末端精确对位的动作
不适合的情况	手必须停在某处不动——你会发现怎么调都差一点	末端的目标位置本身就在大幅运动——反解会"发抖"或超范围
常见工具	直接旋转骨骼（多数软件默认）	求解器：3ds Max 的 HI / HD / IK Limb Solver；Blender 的 IK 约束；UE5 的 IK 节点
注意事项	逐级累积，很难"一眼定位"末端到达点	可能无解（目标太远）或多解（肘朝上/朝下）——需要额外控制；**不要过度使用 IK**
导出的坑	——	导出前通常要执行**烘焙动画**（Bake Animation），把约束与 IK 解算结果固化成逐帧的 FK 数据，否则引擎无法正确重现

∫ 深入一格 · 两段骨骼的 IK 反解：一道中学几何题 选学 · 数学

把一条手臂简化为两根直棒：上臂长 L_1、前臂长 L_2，肩在原点，手要到达距离肩为 d 的目标点。软件是怎么反解的？答案是一百多年前就写进教科书的**余弦定理**：

$$\cos\theta = (L_1^2 + L_2^2 - d^2) / (2 \cdot L_1 \cdot L_2) \quad ，其中\ \theta\ 是肘部的夹角$$

θ 为三角形在"肘"顶点的内角。由肩、肘、手三点围成的三角形中，边长分别是 L_1、L_2、d。

过程就是：由 d 与两边长求出肘的夹角 θ，再求出整个手臂的朝向，两个关节的角度就都定下来了。这个式子还顺手解释了 IK 的两个"脾气"：

- **够不到就无解。**三角形的两边之和必须不小于第三边（$L_1 + L_2 \geq d$）。若目标点比手臂还远，式子里的括号值会超出余弦的取值范围——软件只能"放弃"：要么伸直手臂"指向"目标（明显够不到），要么拉伸骨骼（形变失真）。这就解释了为什么把脚"钉"在地上的时候，腿长的角色一旦比例对不上就会露馅。
- **够得到也未必唯一。**同样的三角形，肘可以朝上弯，也可以朝下弯——数学上有两个解。所以专业的 IK 设置里都配有一个"极向量/控制点"（pole vector）来告诉软件"肘朝这边"。动画里这个手肘方向的瞬间翻转，观众一眼就能看出来，是常见的翻车点之一。

最后回到实践：IK 主要是**动画时的工具**，不是**交付时的数据**。绝大多数游戏引擎与下游工具期望拿到的是纯粹的关键帧数据，因此在导出前把 IK"烘焙"成 FK 是标准工序。你在网页工具里看不到这两个词，但只要它提供"把手拖到某处"或"脚部锁定"，背后都是同一件事。

5.6 动画十二原则：给眼睛配一把尺

现在你已经能让角色动起来，也能让它动得"技术正确"。但技术正确和"好看"之间还有一道鸿沟，跨过这道鸿沟的通行证，是 1981 年由两位迪士尼动画师 **Ollie Johnston 与 Frank Thomas** 在《生命的幻象》（The Illusion of Life: Disney Animation）中总结的**动画十二原则**。这十二条最初来自手绘二维动画的经验，但它们描述的是"人类大脑如何读懂运动"——因此对三维动画、对 AI 生成的动作，一样适用。

本书不对十二条平均用力。作为数字媒体创作者，你需要的是把其中一部分变成**质检员的眼睛**：在检查一段 AI 生成的动作时，能一眼看出"哪里不对劲"。下面先给全貌，再逐条讲与检查最相关的几条。

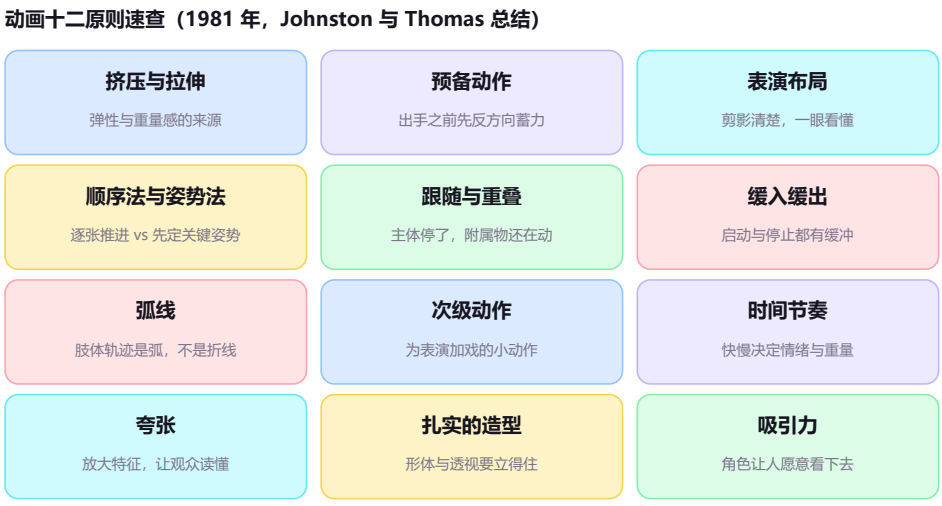


图 5-7　动画十二原则速查。十二条一起背没有意义。建议先记住与动作质检关系最紧的五条——预备动作、跟随与重叠、缓入缓出、弧线、时间节奏，其余各条在后续创作中对照查阅即可。

◆ 先讲两条最"救命"的：预备动作与跟随重叠

如果一段 AI 生成的动作让你觉得"哪里怪但说不上来"，十有八九是这两条缺失导致的：

预备动作（anticipation）：在做一个大动作之前，先向反方向做一个小动作。跳起来之前先下蹲、出拳之前先收拳、举杯之前先把身体微微后仰。它的作用有两个：物理上，它是蓄力；心理上，**它给观众一个"预告"，让观众提前半拍做好准备。**没有预备动作的动作，在观众眼里就像"凭空发生的"——不是不自然，而是"没有重量"。检查法：放慢速度看动作起点，问自己"它出手之前，有没有那一瞬间的反向蓄力？"

跟随与重叠（follow through and overlapping action）：主体停下来之后，附属物（衣袖、头发、飘带、末梢）还要"余势"继续运动一段，然后才慢慢停住；同时，身上各部分的运动在时间上有先有后——上身先动、衣摆后跟。这一条的物理根源是惯性：**轻的东西比重的东西"反应慢、停得晚"**。在三维里，它对应骨骼链末端的延迟旋转、以及附加物理模拟。检查法："角色的手停下来之后，袖口是不是'啪'地立刻静止了？"——是的话，这段动作就缺一口气。

◆ 其余相关条目：从"看得到"到"看不出"

弧线（arcs）：人体的关节做旋转运动，因此手、脚、头的运动轨迹天然是弧线。AI 生成的动作若出现折线式的轨迹（比如手在空中"拐直角"），多半是插值或姿态重建出了问题。检查法：在软件里打开运动轨迹显示，或者录屏后逐帧看手的路径。这条与下一节的重定向问题直接相关。

时间节奏（timing）：同一个动作，用 8 帧做完和用 24 帧做完，传达的情绪完全不同——快等于果断、紧张；慢等于沉重、迟疑。这也是 5.4 节曲线的作用所在。检查法：蒙住画面只看"每一段各占多少帧"，把数字写下来，问自己"这个时间分配讲得通吗？"

次级动作（secondary action）：为表演加戏的小动作——说话时的手势、呼吸时的肩膀起伏。它与主要动作的关系是"烘托而非抢夺"：好的次级动作让主表演更清楚，坏的次级动作会分散注意力。检查法："我的眼睛有没有被不该动的地方吸引走？"

夸张（exaggeration）：把特征放大到超过真实，让观众读得懂。注意夸张要"有分寸、有一致性"：卡通角色可以压得很扁，写实角色夸张的是分寸与节奏；同一个角色的夸张程度要在全片保持一致，忽大忽小会显得失控。**挤压与拉伸**是最典型的夸张手段，但它是软组织角色的特权——硬表面角色（机器人、家具）上用挤压拉伸，效果反而是灾难。

把十二原则里剩下的几条一次说完，让"速查图"变成真正能用的知识。**表演布局**（staging）关心的是"观众能不能看清"：同样的动作，从哪个角度看、剪影清不清楚、有没有被杂物遮挡。三维里它的对应检查非常干脆——定格截图，眯起眼睛看剪影：如果你自己都要猜"这只手在干嘛"，观众更看不清。**顺序法与姿势法**（straight ahead / pose to pose）是两条制作路线：前者从头到尾逐帧推进，自由但难控制；后者先摆关键姿势再补中间，可控、适合叙事性动作。三维的主流工作流（Blocking 起步）就是姿势法。**扎实的造型**（solid drawing）在手绘时代指形体、比例与透视要立得住；到了三维，它被"继承"成了本章前半部分的工作——模型合格、权重干净、形变不塌，就是三维版本的"造型扎实"。**吸引力**（appeal）则不是一项独立技术，而是其他十一条到位之后的副产品：观众愿意看完第一遍、还想看第二遍，就说明它存在。检查法朴素得近乎残酷：放给同伴看，观察他的眼睛有没有走神。

表 5-7　与"检查动画质量"最相关的原则：三维实现与检验问题

原则	三维里的实现手段（来自业界常用做法）	检查时问自己
挤压与拉伸	晶格变形器、混合形状（morph）、骨架上缩放特定部位	软的东西该变形的地方有变形吗？硬的东西是不是被错误地压扁了？
预备动作	在关键帧序列前段插入反方向的关键姿势；用 FK 摆出蓄力姿态	动作起步前一瞬间，有没有反向蓄力？
跟随与重叠	骨骼链末端的延迟旋转、附加物理模拟（毛发、衣摆）	主体停下后，末端有没有"余势"再停？
缓入缓出	曲线编辑器里的切线调节（自动/钳制/平滑等）	有没有哪一段在匀速运动？发力点是否干脆、收势是否柔和？
弧线	观察运动轨迹；用 IK 时检查末端路径；修插值曲线	手、脚的轨迹是弧线还是折线？
次级动作	独立控制的表情、手势、衣摆骨骼层	小动作是"加戏"还是"抢戏"？
时间节奏	摄影表整体调帧距；曲线局部的快慢对比	把每段的帧数写下来，这个分配符合你要的情绪吗？
夸张	幅度与停顿的放大；禁止在硬表面角色上滥用变形	夸张的程度在整段里一致吗？有没有超出角色允许的形变范围？

> 💡 **十二原则与"AI 生成动作"的关系**
>
> 有人会问：这些是手绘时代的经验，对 AI 生成的动作还有意义吗？**恰恰更有意义。**因为 AI 生成动作时，学到的是"数据中出现频率最高的运动模式"——它能生成正确的走路、正确的挥手，但它不会主动替你想"这个角色此刻的表演该怎么设计"。预备动作、次级动作、夸张，本质上是**表演决策**而不是**运动规律**；决策必须由人来做。这解释了本章的一句常挂嘴边的话：AI 动作给你一个"会动的角色"，把它变成"会演戏的角色"，是你的工作。

5.7 运动条件与动作重定向：把动作"翻译"到另一个角色身上

> **📌 本节与第 6 章的边界：原理版 vs 实操版**
>
> 动作重定向在两章里各出现一次，分工是这样切的：**本节只回答"为什么能搬、什么条件下搬得动"**——四类前置条件、翻译类比、以及"旋转可搬、位置要换算"这条量纲直觉；**第 6 章 §6.5 回答"具体怎么搬"**——骨架映射、比例适配、足部锁定这三道关，以及导出落地与失败排查。已经读过第 6 章的同学，可以把本节当作它的"原理注脚"；还没读的，请把它当成上路前的执照。

到这里，本章的技术部分只差最后一块，也是本章牵向下一章的接口：**你拿到的动作数据，往往不是你这个角色的。**Mixamo 的走路、CMU 数据库中真人的跑步、HY-Motion 生成的"举杯"——这些动作都长在别的骨架上。把它们迁移到你的角色身上，这个过程叫**动作重定向**（motion retargeting）。

先回答一个前置问题：**一段动作数据，要满足什么条件才能顺利地被套用？**这就是"运动条件"：

- 条件一 **骨架命名规范对得上。**动画数据本质上是"每根骨骼在每个时刻的旋转"。如果源骨架叫你 `LeftForeArm`、目标骨架叫 `forearm_L`，软件不知道它们是一回事，重定向就无从谈起。这就是为什么 Mixamo 兼容的骨骼命名（Tripo、Meshy 等工具都采用）有工程价值——它相当于给跨工具重定向装了一个"通用插头"。
- 条件二 **比例相似度够高。**命名解决"谁是谁"，比例解决"换个尺寸是否还成立"。同样 45 度的髋部旋转，放在腿长 0.9 米的角色上脚刚好落地，放在腿长 0.7 米的角色上就够不到地。源骨架与目标骨架的比例差距越大，需要人工修正的地方就越多。
- 条件三 **帧率与时间基准一致。**源动作是 30 fps 还是 24 fps、有没有做过时空缩放，都直接影响套用后的速度感。多数工具会自动处理，但处理错时症状很怪：动作看起来"快半拍"或"像开了慢放"。
- 条件四 **根运动（root motion）的约定相符。**动作数据里是否包含"角色整体位移"？有的动作原地蹬腿（in-place，位移交给程序控制），有的自带位移轨迹（root motion）。混用两者，会出现"原地踏步却位置漂移"或"该前进却在原地跑"。

四个条件满足得越好，重定向就越像"即插即用"。而重定向本身，可以用一个精确的类比来理解：它像翻译——把一句用"源骨架"这门语言写成的动作，翻译成"目标骨架"的语言。翻译的四个步骤，对应重定向的四个环节：

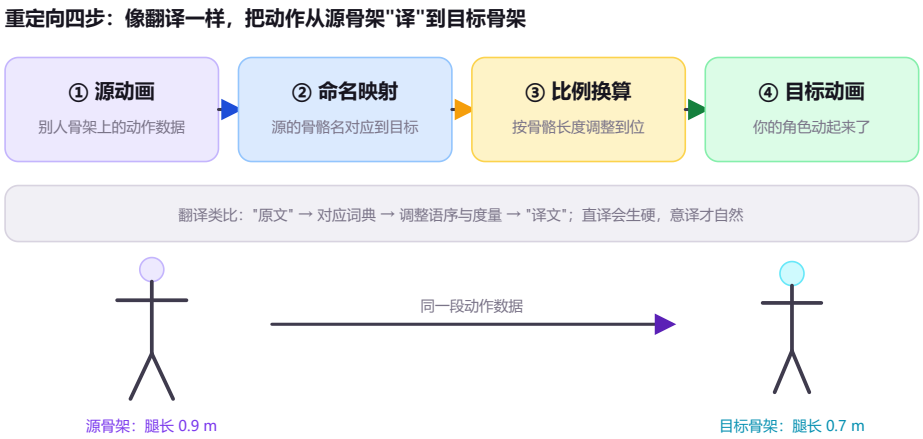


图 5-8　重定向四步。命名映射决定"能不能对上"，比例换算决定"对不对得上"。四步之中，人工工作量最大、也最容易翻车的是第三步——它正是下面"滑步"问题的源头。

∫ 深入一格 · 什么能直接搬，什么必须换算 选学 · 原理

重定向最反直觉的一点是：**旋转可以直接搬，位置必须换算。**为什么？

因为"人体怎么转动关节"这件事，与身高无关。一个人跨步时髋部转 30 度、膝盖转 20 度，这个"姿态描述"在高个子和矮个子身上是同一套语言——旋转描述的是"朝向的变化"，是一个无量纲的量，所以只要骨架的拓扑（哪根骨骼连着哪根）对应得上，角度的数值可以直接平移过去。

而位置是有量纲的。骨骼长度不同，同样的角度会把末端送到不同的地方——这不是软件"没算对"，而是几何上必然如此。必须按比例重新换算的东西包括：**骨盆的高度与起伏**（决定腿是伸直还是弯曲）、**步幅与前进速度**（决定脚踩在哪里）、以及所有以世界坐标表达的目标（脚的目标点、手抓取的位置）。

做一个最小化的算例：源角色腿长 0.9 米、一步跨 0.7 米；目标角色腿长 0.7 米。若照搬同样的髋部与膝盖角度，目标角色的脚只能落到约 0.55 米处——每步比"应该在的位置"短了 0.15 米。在屏幕上，观众看到的不是"腿短"，而是**脚踩在地面上的那段时间里还在向前蹭**：这就是滑步。修它的办法，正是"意译"：保持动作的**感觉**（节奏、幅度关系）不变，把**落点与位移**重新定标。

同一个髋角，两副腿：长的踩稳了，短的够不着

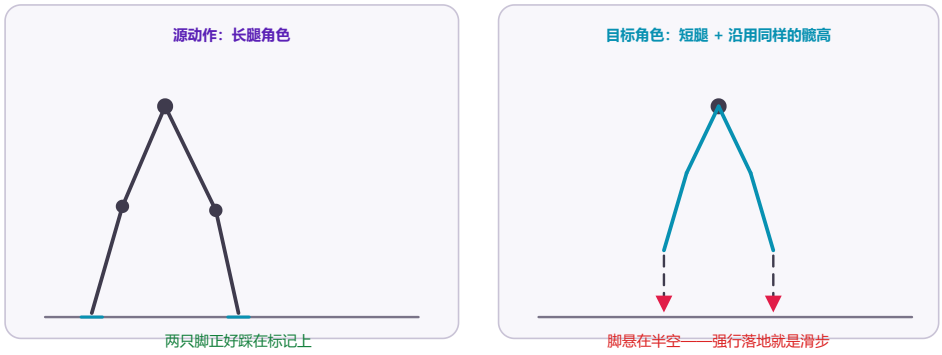


图 5-9　比例差异与滑步。滑步的根源不是"AI 不聪明"，而是几何：骨骼长度变了，同样的角度就送不到同一个位置。重定向的第三步（比例换算）做得好不好，观众用眼睛就能投票。

🩺 失败诊所 F5-2　踩在冰面上的诗人

症状：小舟从动作库给诗人角色套了一段"走路"：从背后看挺像回事——摆臂、迈步都齐全。可一放到地面上，脚底像踩在冰面上滑；切到侧视再细看，支撑脚落地后还在缓缓前移。换了一段"拱手作揖"，问题更严重：手腕直接穿进了胸口，衣袖从肋部"长"了出来。

成因：两个问题同源——**源骨架与目标角色的比例不匹配。**动作库里的走路来自一个腿长、髋宽都是"标准成人"的角色；诗人角色是从第 4 章的三维生成流程来的，比例有自己的脾气（略瘦、臂展短一点）。于是：① 步幅照搬，但腿长不同 → 落点对不上，脚在地面上滑动；② 手臂弯曲的姿势照搬，但臂展与胸廓厚度不同 → 手腕的路径切进了胸腔。注意，两个骨架的命名是兼容的（条件一满足），所以才"套得上"——这恰恰说明：**套得上不等于套得对。**

处方：按"先动因、后症状"的顺序处理。① **修步幅**：在重定向设置里按腿长比例缩放位移与步幅；若工具不提供，手动把"脚落地"的关键帧向前挪，让支撑脚在整个支撑相里保持不动（脚锁定）。② **修路径**：把穿模的手腕关键帧向外微调，或对穿模段使用 IK 把手腕约束在体表之外，并检查手肘的极向量方向。③ **修节奏**：比例调整后动作的时长感会变，回到曲线视图检查关键帧间距。④ **兜底判断**：如果一段动作库的体型假设与你的角色差异过大（比如巨人动作给小个子），最好的处方可能不是修——而是换一段动作，或者自 K 关键帧。

经验值：行业实践中，重定向后"完全正确"的比例并不高；把"修好滑步与穿模"当作套用动作库后的**必经工序**，而不是意外事故——心态对了，返工就不痛苦。

重定向的「操作版」——三道关怎么过、脚怎么锁住、导出前查什么——见 第 6 章 §6.5；那里还会给出「不想自己拍」的三条现成动作路线。

表 5-8　重定向常见问题与对策（按症状查）

症状	背后的原因	对策
脚在支撑相滑动（滑步）	步幅/位移按源腿长，未按目标腿长定标	按腿长比例缩放位移；用脚锁定把支撑脚钉住；重排落地关键帧
脚陷进地面或悬空	骨盆高度照搬源数据	调整根骨骼高度；对双脚使用 IK 约束到地面；检查脚掌的滚动
肢体穿模（手穿胸口、臂穿衣摆）	臂展/躯干厚度不同，路径直接照搬	外移关键帧；用 IK 约束末端；加"最小安全距离"检查
动作幅度看起来"缩水"或"过冲"	两侧比例不同导致同样的角度产生不同观感	按需整体缩放旋转幅度；回到曲线视图修节奏
动作"快半拍"或像慢放	帧率基准不一致，或时空缩放被误用	统一时间基准（如 30 fps）；按实际时长重新校核关键帧
原地踏步却位置漂移，或该前进却在原地跑	根运动约定不符（in-place 与 root motion 混用）	确认下游要哪种约定；导出时选择匹配的选项；该去掉的位移要去干净
手肘/膝盖瞬间"翻向"另一侧	IK 多解跳变，极向量缺失或目标越界	补极向量控制；检查目标点是否超出可达范围（$L_1+L_2 \geq d$）
循环动作在接缝处"抽一下"	首尾帧姿态与曲线不匹配；数据未做循环裁剪	对齐首尾关键帧；把循环外插设置为 cycle；必要时手动重排接缝处的两三个关键帧

5.8 AI 工具现状：绑定、文生动作与动作库

本章前面所有原理，此刻都落到一个实际问题上：**这些工作，今天有哪些工具能替我做、替到什么程度？**答案分成三块：自动绑定（让骨架长出来）、文生动作（让动作长出来）、动作库与公开动捕数据（让动作"借"过来）。这三块恰好覆盖了传统管线里最耗时的环节，也正是行业对 AI 期望最集中的位置。

◆ 5.8.1 自动绑定：把手工业变成"秒级 + 验收"

回忆 5.3 节的场景：手工绑定的流程是"重拓扑 → 手动逐关节放置骨骼 → 逐顶点刷权重 → 摆极限姿势测试"，一个熟练的绑定师也往往要花数小时到数天。自动绑定工具把中间两段变成了秒级到分钟级的自动流程：上传网格（T-pose 或 A-pose）→ 用计算机视觉检测关节位置（先进一些的工具还会输出每个关节的置信度）→ 拟合一套标准骨架 → 计算蒙皮权重 → 导出带骨架与权重的 FBX/GLB。这正对应本章的输入条件——那些标准骨架几乎是同一个规格：例如 **Kinetiq Engine** 拟合的骨架含 Foot→Ball→Toe 的完整脚链，可选逐指手部检测；**auto-rig-web** 用 ONNX 视觉模型 + WebGPU 在浏览器本地跑出标准 19 关节双足层级；**AccuRig 2.0** 提供 19 关节身体骨架加独立手指骨架。

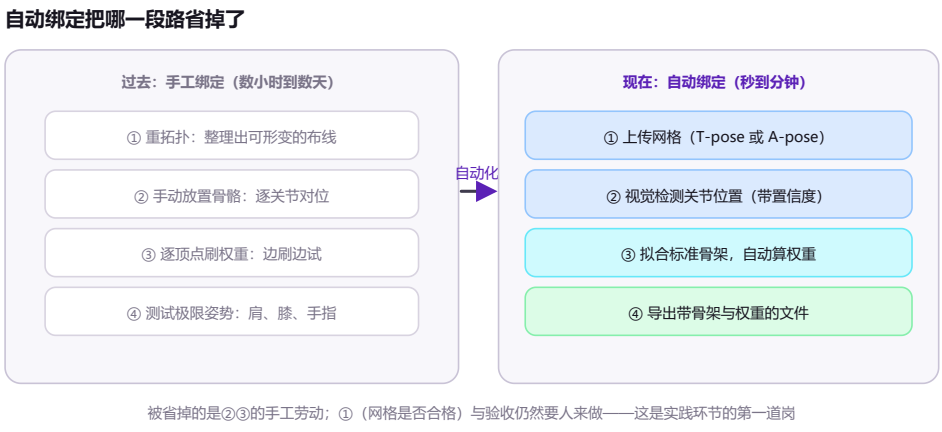


图 5-10　手工绑定与自动绑定的流程对比。自动绑定不改变"绑定"这件事的本质，只是把其中重复性最强的两段自动化了。因此关于权重的判断力（5.3 节的检查清单）依然是你的工作——而且是不可外包的那部分。

下表汇总了检索到的自动绑定工具。需要特别说明来源性质：关于 **Mixamo** 现状的对比信息主要来自 Meshy 的博客——它们是竞品的口径，教材引用时如实标注。

表 5-9　自动绑定工具对照（工具能力与额度以官方页面为准）

工具	形态	能力要点	免费额度 / 限制
Mixamo（Adobe）	网页	免费的浏览器端标准双足人形最快路径；**约 2,500 个免费动捕片段**	据竞品（Meshy）对比：2015 年 Adobe 收购后进展停滞，2025 年 6 月曾发生多日宕机；商业许可条款需细读
Meshy	网页	浏览器端自动绑定；人形、四足与智能检测；构建完整骨骼层级；**约 30 秒**计算蒙皮权重；**Mixamo 兼容骨骼命名**；600+ 动画预设；导出带骨架与权重的 FBX/GLB	免费 100 credits/月，无需信用卡；**免费输出标 CC BY 4.0（需署名）**
Tripo Studio	网页	一站式工作区：生成角色并原地自动绑定（T-pose，"Mixamo-ready"）；**UniRig** 一键绑定含完整骨架与蒙皮权重，覆盖双足/四足/多足；导出 GLTF/GLB、FBX	免费 200 credits/月；**条款保留宽泛权利**——商用角色建议付费档或书面确认
Kinetiq Engine	网页（客户端）	一键 Auto-Rig；关节检测给出置信度；骨架含 Foot→Ball→Toe 完整链；可选逐指手部；**"no-bleed" 蒙皮权重**；可重定向 Mixamo/VRM 文件；导出 FBX/GLB/BVH 等	绑定与预览免费；导出需 Starter 档；v1.1 于 2026 年 6 月发布
auto-rig-web	开源（浏览器本地）	零点击绑定：上传静态 gltf 即得 IK-ready 蒙皮网格（毫秒级）；标准 19 关节双足层级；程序化蒙皮权重；含 CCDIK 求解器	开源免费；无需服务器、无需上传
Kinemo	网页	面向**非人形**（龙、机器人、六足怪物）：通用自动绑定、文生动作、骨架/关键帧编辑器	有免费 credits
AccuRig 2.0	桌面（免费）	19 关节身体骨架 + 独立手指骨架；导入 OBJ/FBX，导出 FBX/USD，含主流引擎预设；可浏览/重定向库存动画	需免费账号才能导出；免费版限单网格；面部绑定与权重编辑需付费工具
Rigify	Blender 插件	基于 meta-rig 手工搭骨架，含 FK/IK 控制	免费；**不是自动检测工具**——需要手工搭建与动画，适合学习骨架结构

给本章实践的一条结论：纯网页、免费、且能完成"上传网格 → 得到带权重的骨架"这条链的候选，是 Meshy、Tripo Studio、Kinetiq Engine 与开源的 auto-rig-web；但它们都不能完全替代 Mixamo 的免费动捕库——绑定能力可以比 Mixamo 更强，动作库不行。所以课程任务采用"组合拳"：好一点的绑定用新工具，找动作先去 Mixamo。另外注意许可证差异：免费档常带 CC BY 4.0 署名要求，条款也可能"保留宽泛权利"——课程作业引用时按第 1 章"三查"的习惯记录来源与授权；商用角色则务必付费或书面确认。

◆ 5.8.2 文生动作：用一句话生成动作

如果说自动绑定解决"骨架从哪来"，那**文生动作**（text-to-motion）解决的是"动作从哪来"。你输入一句人话——"缓慢抬起右手举杯，抬头看月亮"——模型输出一段可以直接用在角色上的骨骼动画。这个方向信息最完整的中文生态模型，是腾讯混元的 **HY-Motion 1.0**。

图 5-11　文生动作管线。从"文字"到"骨骼动画"之间站着三层：LLM 把模糊指令翻译成结构化描述与时长（"智能动作导演"）、扩散 Transformer 与流匹配生成动作、再经标准骨架输出给下游。你不必懂每层的实现，但要懂它的**输入契约**：写得具体、写得短。

表 5-10　HY-Motion 1.0 关键事实（截至本书成稿；细节以官方发布为准）

项目	内容
是什么	文生 3D 人体动作生成模型，基于扩散 Transformer（DiT）+ 流匹配；官方称是首个把文生动作领域的 DiT 扩展到十亿参数级别的模型
发布节奏	2025-12-30 发布推理代码与预训练模型（HuggingFace / 官网）；2026-01-08 多家媒体报道开源；2026-01-29 发布 SSAE 评测（一种基于视觉语言模型评估生成动作语义对齐度的方法）
模型库与硬件	HY-Motion-1.0：1.0B 参数，最低约 26 GB 显存；HY-Motion-1.0-Lite：0.46B，最低约 24 GB 显存
三阶段训练	① 3000+ 小时多样化动作数据大规模预训练 → ② 约 400 小时精选高质量 3D 动作数据微调 → ③ 基于人类反馈与奖励模型的强化学习
数据管线	融合单目视频动捕、光学动捕与艺术家手 K 动画，统一重定向到标准骨骼、清洗为 30 fps 切片；标注走"渲染 → 视觉模型初标 → 人工校验 → 大模型结构化扩写"的闭环
动作覆盖	基础移动、日常生活、社交休闲、健身户外、体育竞技、游戏角色动作——六大领域、200+ 细分动作类别
"智能动作导演"	LLM 模块负责改写用户指令并估计动作时长，把模糊描述转成"结构化英文描述 + 精确时长"
技术规格	201 维动作特征、30 fps；文本编码器为 Qwen3-8B + CLIP-L 句嵌入；Euler ODE、50 步采样、CFG scale 5.0
接口与集成	支持 macOS / Windows / Linux；提供命令行批量推理与 Gradio 界面；生成的动画可接入 Unreal Engine、Unity、Maya、Blender
明确不支持（官方）	非人形角色；情绪、服饰与外观属性；场景或镜头角度；多人动作；循环或原地动画
许可证	tencent-hunyuan-community（HuggingFace 标为 license: other）——使用前读条款原文
社区生态	已有社区工具包可把生成动作经 SMPL ↔ MMD 重定向导出为 VMD（用于 MikuMikuDance 一类的应用）

把上面这张表读成一句实践建议：文生动作适合生成"单人的、不带位移的、几秒钟的"表演片段——这恰好就是本章主线任务（举杯邀月）的形状。而"循环的走路""多人对打""带情绪的表演"不在它的能力圈里，需要换工具或换思路。

✎ 提示词 • 文生动作（主线任务「举杯邀月」）

【第一步：中文意图，先写对，再写短】
一位古装诗人，站立，缓慢抬起右手把酒杯举到胸前，抬头望向天空，停顿片刻。

【第二步：按"结构化英文描述 + 精确时长"改写】
A poet stands still, slowly raises his right hand lifting a wine cup to chest level, then tilts his head up to gaze at the sky, holding the pose briefly.

约束：
- 只描述动作本身——**不写情绪、服饰、场景、镜头**（模型不支持，写了也是噪音）；
- 时长：动作总长 **4–5 秒**；提示词尽量短（官方建议少于 30 词）；
- 只生成单人、原地动作。

输出：可直接接入角色的骨骼动画（30 fps）。

工具 HY-Motion 1.0　时长 4 秒　种子 1（省显存）　关键 只写动作，不写修饰

◆ 5.8.3 动作库与公开动捕数据：不生成，直接"借"

第三块资源是"现成的动作"：预录好的动捕片段库。它比文生动作老得多，但至今不可替代——因为动捕数据是"真人的真实运动"，自带全部微妙的重量感与习惯动作。教材的任务是教会你**挑库、读许可、知道数据要加工。**

表 5-11　公开动捕数据与动作库：规模、许可与使用代价

资源	规模	格式	许可与使用要点
CMU MoCap Database	2,605 个 trial，6 大类 23 子类（走路、体育、日常活动、交互）	原始 ASF/AMC；广泛以 BVH 分发	**免费用于所有用途**（含商业产品），**但不可直接转售**（转换格式后也不行）；需按规范致谢
Mixamo（Adobe）	几百个动画（移动、战斗、交互）	引擎兼容的骨架动画	Adobe CC 条款，商业许可需细读；**预解算、循环就绪、引擎兼容**，适合原型与占位符
cgspeed 转换版	同 CMU	转 3ds Max / MotionBuilder 的 BVH	许可同 CMU；免去手工转换
HuggingFace CMU 动捕库（FBX 版）	2,548 个 FBX	FBX	许可同 CMU；**仅骨骼无网格、含根运动、无手指动画**
GitHub CMUMocap	CMU 数据的 Unity 版	Unity humanoid 动画	许可同 CMU；可直接用于 Unity

读这张表要读出"代价"。CMU 数据的许可堪称慷慨（免费、可商用、可修改，只要不转售），但它的数据是**研究器材的原始输出**：无引擎骨骼绑定、无循环裁剪、无脚部固定清理——17 关节层级且命名非标准，脚趾与手部关节有噪声（手指与拇指数据实际未采集）。把它变成可用的产品素材，要经过 BVH→FBX 转换、重定向到引擎骨架、修脚部滑动、裁剪循环点这四道加工。这就是"运动条件"（5.7 节）的现实版本：动作数据的许可成本很低，加工成本才是大头。而 Mixamo 恰好相反：动捕出身、预解算、循环就绪、命名标准，代价是覆盖面薄（标准移动与基础战斗之外内容少）且多年未更新。

◆ 5.8.4 选路：你的任务该用哪块拼图

三块拼图各有脾气，选错了工具，后面所有时间都在还债。给一张最简的决策树，遇到新任务时按顺序问三个问题：

第一问：角色是人形吗？不是——龙、机器人、六足怪物这类非人形角色，文生动作明确不支持（模型是按人体骨架训练的），需要找面向非人形的绑定与动画工具，或在三维软件里手 K。是——进入第二问。

第二问：动作需要"循环"或"带位移"吗？走路、跑步、待机这类要无缝循环的移动，或者要精确控制前进距离的位移动作——文生动作不支持循环与原地动画，动作库才是本命（库里的片段本身就是"循环就绪"的）。而"单人的、几秒钟的、原地的小表演"（招手、鼓掌、举杯）恰好是文生动作的主场。

第三问：动作要和环境/道具精确交互吗？要——手扶栏杆、脚踩台阶、按住按钮，人工参与的成分最高：要么用动作库里带交互的片段再精修，要么直接 K 关键帧，用 IK 把接触点"钉"住。不要——回到第二问的结论即可。

三个问题的答案组合起来，就得到本章实践环节的标准路线：人形 + 单人短表演 + 无环境交互 = 文生动作；人形 + 循环移动 = 动作库；带交互或非人形 = 人工为主。把这棵决策树抄进你的创作档案——它比"哪个工具最强"这个问题有用一百倍。

> 📊 **案例 5-C** § "2 分钟、1 美元"的另一面
>
> 工具厂商们乐于宣传的量级数字：Meshy 官方口径称"过去需数周、花费 1000 美元的 3D 资产，现在 2 分钟、1 美元"；其上层的自动绑定约 30 秒计算蒙皮权重；Tripo 的 Smart Mesh 号称约 2 秒出游戏就绪拓扑。**请务必带着"来源性质"读这些数字**——它们是企业自称或券商测算，口径各异（算不算人工时间？算不算失败重试？算不算后续修复？）。
>
> 本课程希望你们记住的是这些数字**没有**包含的部分：一次自动绑定后，验收权重、修渗漏、修滑步、修穿模、对齐口型与道具的时间——本章的失败诊所与动手实践，花的主要就是这笔时间。工具的贡献是把"从 0 到 1"压缩到分钟级；"从 1 到能交付"这一段，目前依然是人的工作时间，而且是最需要审美判断的那一段。

5.9 材质与缺陷修复：把角色"体检"合格再动起来

现在请把视线从"动作"暂时移回"模型"。因为它们马上要合体了——而在合体之前，有一个残酷的经验事实：

> ⚠️ **一条被反复验证的行业经验**
>
> 动画会放大一切问题。静态时看不太出的接缝、拉伸与错位，一进入动画就会"活"给你看：肩部的图案在抬臂时被扯长，腰带的边界在转身时裂开。一句话——静态检查通不过的模型，动起来一定更糟。所以专业管线里，"绑定前体检"是一门必修课：网格、UV、贴图、法线，逐项过一遍，再交给绑定。

◆ 5.9.1 四块拼图：UV、法线贴图、PBR、色彩空间

第 4 章已经说明模型从哪来，本节补上"贴图怎么贴得对"。四块拼图分别是：

UV：把三维表面"摊平"到二维图上的展开图，相当于给角色做了一份"服装裁剪图"。所有贴图（颜色、细节、粗糙度）都要按这份裁剪图对应上去。UV 有两个关键工程量：**间距**（UV 岛之间要留出空隙，防止贴图过滤时"串味"——1K 贴图至少留 4 px，2K 至少留 8 px；烘焙用的贴图还要更大的 margin，工程上常取 16 px 以上）与**纹素密度**（贴图像素与实际表面的比例要全模型一致，否则脸上精细、手上模糊）。坏 UV 造成的拉伸与接缝，没有任何提示词能修。

法线贴图：用一张贴图记录"表面的朝向微变化"，让低多边形模型看起来有高模的细节。它是"看起来的凹凸"，不是真的把几何捏出来。两个实践要点：一是**方向约定**——不同软件/引擎对绿色通道的方向约定不同（OpenGL 约定与 DirectX 约定相反），用错约定的症状很典型："凹陷看起来像凸起"；二是**别对它做 AI 去噪或重生成**——法线贴图每个像素都必须是长度为 1 的向量，被"美化"过就废了。

PBR：一套"基于物理"的材质描述体系，把材质拆成几张各管一事的贴图：底色（反照率）、法线、粗糙度、金属度、环境光遮蔽等。它的好处是材质在各种光照下都"讲道理"，是当代游戏与实时渲染的通行标准。各通道的"可靠度"在 AI 生成管线里差异极大，下表给出教材口径。

色彩空间：贴图不是"图"，是"数据"。底色（反照率）与发光是给人看的颜色，按 sRGB 处理；法线、粗糙度、金属度、AO 是数值数据，必须按 Non-Color（线性）处理。把数据贴图当颜色贴图导入，粗糙度与深度就会整体偏掉——这是新手最常见的"材质发灰/发脏"原因之一。

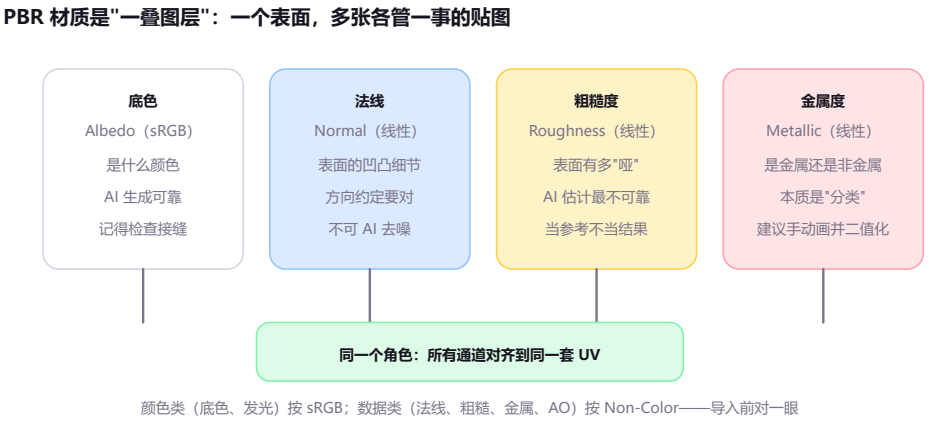


图 5-12　PBR 材质的分层结构。训练自己的一个习惯动作：拿到 AI 生成的贴图，先看"这套 UV 是否可用、通道是否齐全、色彩空间是否标对"，再谈好看——顺序反了，后面所有打磨都是白工。

表 5-12　PBR 通道：管什么、什么色彩空间、AI 生成中的可靠度

通道	它控制什么	色彩空间	AI 生成管线中的可靠度（工程经验）
底色 Albedo	物体本身的颜色与图案（不含光照）	sRGB	相对最可靠，主要风险是多视角接缝与"光照被烘进颜色"（底下再打光会双重变暗）
法线 Normal	表面朝向的细节起伏	Non-Color（线性）	中等：结构常对，但方向约定与强度需人工核验；**不能做 AI 去噪**
粗糙度 Roughness	表面反光的"锐利/模糊"程度	Non-Color（线性）	**最差**：生成值常整体偏差，只能当"参考"，几乎都要手工重画关键区域
金属度 Metallic	该处是金属还是非金属	Non-Color（线性）	**很差**：这是"分类"而不是"回归"问题，AI 常生成模棱两可的中间值——建议手工分区域画成接近二值的黑白
环境光遮蔽 AO	环境光被遮挡的暗处（如腋下、折角）	Non-Color（线性）	多由几何烘焙而来，与 AI 生成无关；注意不要与"烘焙进底色的阴影"重复叠加
高度 Height	供重新烘焙法线/置换用的高度信息	Non-Color（线性）	中等：可用于细节迁移，但直接当置换会撑爆面数

🎯 引擎间的"翻译表"

不同引擎对通道的打包方式不同，认得出这几种常见变体就不慌：glTF 系将金属度与粗糙度打包在同一张图的蓝色与绿色通道；Unity 的"光滑度" = 1 − 粗糙度（数值相反，意思相同）；Unreal 常把 AO、粗糙度、金属度打包在 R/G/B 三通道。**本质上是同一套信息的不同装订方式**——遇到"贴图丢通道"的问题，先怀疑打包方式，而不是怀疑 AI。

◆ 5.9.2 常见生成缺陷与"修复优先级"

接下来是本节的核心工具：一张"缺陷—症状—病因—处方"的对照表，以及一个决策顺序——修复优先级。先记住决策原则：**从"根"修到"叶"**。网格（几何）是根，UV 是干，贴图是叶：叶子上的问题，常常是树干注定的。

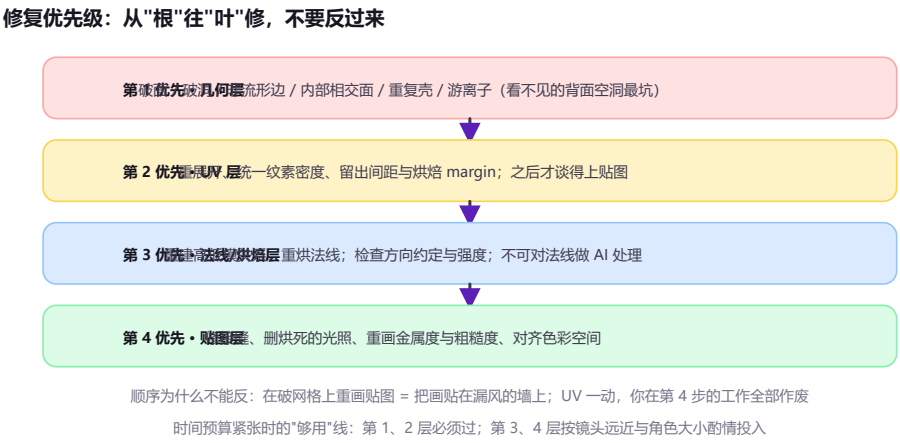


图 5-13　修复优先级的决策链。把这张图当检查表用：每修一个缺陷，先问"它属于哪一层"，再去对应的层动手。跳过第 1、2 层直接修贴图，是新手最常见的时间黑洞。

表 5-13　常见生成缺陷：症状、病因与处方（按修复优先级排列）

层	缺陷	典型症状	病因与处方
几何	**破面与破洞**	正面看完全正常，转过去背面藏着破洞；或表面出现"裂缝"	单视图重建必须"幻觉"看不见的部分，背面容易出错。处方：从多角度看一遍模型；用网格修复（填洞、重网格化）工具处理；提供多视角输入可以显著减少此类问题
	塌陷与凹坑	表面出现不该有的凹坑，或局部几何向内塌缩成一团	生成时的几何退化，或减面/重拓扑过度。处方：局部补面重建或回退到上一版；注意**形变区（肩、肘、膝）的塌陷不能靠贴图掩盖**——一旦弯关节就会原形毕露，必须回几何层修
	内部相交面与重复壳	表面看不出来，但绑定后权重乱、切片或物理模拟失败	生成算法在遮挡区留下的内部几何。处方：删除内部面、合并重复壳；导出前跑一次"水密/流形"自检
	面数失控	一个人形角色动辄五十万三角面；文件巨大、绑定卡顿、进不了引擎预算	AI 生成默认追求"看得见的细节"。工程经验：人工清理后约降至十三万面，而游戏可用的角色通常只要几千到一万五千面。处方：重拓扑/减面，并优先保住形变区（肩、肘、膝、面部）的布线
UV	**拉伸与接缝**	纹理上出现被拉长的图案；不同 UV 岛交界处有可见的缝	UV 展开质量不足。处方：重展开；检查"显示拉伸"视图；把高度/法线类贴图烘焙时留足 margin（工程上常取 16 px 以上）
	纹素密度不均	脸上清晰、手上模糊	各 UV 岛比例不一致。处方：按统一标准（如每米 512 像素这一量级）核对；用"棋盘格"贴图目视检查
法线	**方向约定错误**	凹陷看起来像凸起，或者整体"鼓起来"	OpenGL 与 DirectX 的绿通道约定相反。处方：在导入设置里切换约定或翻转绿通道
	"AI 美化"损坏	表面质感诡异、光照不对	对法线贴图做过去噪或重生成——法线像素必须是单位向量，任何"美化"都会破坏它。处方：从几何重烘法线，绝不单独处理法线贴图
贴图	**光照被烘进底色**	模型在自己的场景里发暗、"发脏"，暗部死黑	AI 生成的贴图常自带光照与阴影（相当于"幽灵光源"）。处方：把光照从底色中修掉（重绘暗部，或重新生成无光照版本）——否则和场景光照叠加后变"双重打光"
	金属度/粗糙度崩坏	该亮的金属灰扑扑，该哑的布料锃亮	两个通道本就是 AI 猜得最差的部分。处方：按材质分区手绘：金属度尽量二值化，粗糙度按"金属—布料—皮肤—陶瓷"等类别画，再补细节

🩺 **失败诊所 F5-3** 转身之后，背上有扇"窗"

症状：小月负责给角色贴图。她在正面花了两个小时：皮肤细腻、衣纹清楚、颜色也调顺了——正面堪称完美。然后在场景里转了一下镜头。背面上部有一块两指宽的破洞，边缘还向外翻着；右肩的贴图拉伸得像被扯过的贴纸；腰带的图案在接缝处对不上，错开半厘米。三个缺陷，一起亮相。

成因：三个症状对应三层原因。破洞属于几何层：单视图生成必须"幻觉"背面几何，这块区域正好是生成误差的堆积处——正面看着完全正确，背面藏着破洞，只有转过去看才发现，是单图重建最典型的失败形态。拉伸属于 UV 层：三角形被摊平时被强行拉开，图案自然被扯长。接缝错位属于贴图层：两个 UV 岛分了家，但图案需要跨岛连续——这种问题无法"事后修补"，只能回到 UV 与投影阶段处理，比如在接近实际使用姿态下做投影，而不是在 T-pose 下做完再指望它跟着动。

处方：完整走了一遍修复优先级。① **几何层**：转一圈检查全身（尤其是侧面与背面），把破洞填上、翻转的法线修正；② **UV 层**：重展开，统一纹素密度，留足间距；③ **法线层**：重建后重烘法线；④ **贴图层**：重绘背面受影响的区域，检查跨岛接缝的对齐。**复盘结论（写进了小组笔记）：拿到任何 AI 生成资产，第一分钟做三件事——转一圈看背面、对齐缩放与原点、开背面剔除看有没有"洞"。**这三件事所花的六十秒，相当于后面两个小时的白工保险。

📊 **案例 5-D** § 从"玻璃罐"到"陶瓷罐"：提示词决定可修复性

一个教学级的小实验，来自工具方帮助中心的公开建议：用"玻璃罐"一类提示词去生成三维资产，得到的网格往往需要大量修复——因为算法要同时处理透明、中空与复杂重叠元素；而这些恰恰是它的弱项。把提示词改为**"陶瓷罐"**（实体、简单、不透明）重新生成，成功率会显著提高。

这个实验的教育意义很直接：缺陷修复的第一道工序，其实是在提示词里完成的。当你准备把生成资产用于动画时，先问一句"我能不能把它描述得更'实体'、更简单"——把造型上"炫技"的部分留给后期手工，把生成留给它擅长的部分。同样的逻辑贯穿本章：材料与动作都遵循"先保证结构可靠，再追求细节丰富"。

5.10 关键概念速查：本章的"词汇表"

本章的概念密度是全课程最高的一章。下表把散落各节的核心术语集中起来，供你在实践与复习时随时回查。用法建议：不求一口气背下来，而是**在需要它的时候来查一次**——概念就是这样长进脑子的。

表 5-14　第 5 章关键概念速查（术语 · 一句话定义 · 出处小节）

类别	术语	一句话定义	小节
骨骼与蒙皮	**骨骼 / 骨架**（armature / skeleton）	一组按父子层级组织起来的关节；"armature"是 Blender 对骨架对象的称呼	5.2
	关节自由度与限位	关节能绕哪些轴转动（自由度），以及转动的范围限制（限位）	5.2
	绑定姿态（bind pose）	绑定发生时的姿势，通常是 T-pose 或 A-pose，方便判定归属	5.3
	蒙皮（skinning）	把网格顶点绑定到骨骼上、使网格随骨骼形变的过程	5.3
	权重（weights）	每个顶点受哪些骨骼影响、各占多少比例；权重之和应归一	5.3
	权重渗漏（bleed）	不该管的骨骼分到了别的部位的权重——"牵一发动全身"的元凶	5.3
关键帧与流程	**关键帧**（keyframe）	动画师创建的关键姿势；中间帧由计算机插值生成	5.4
	插值曲线 / 切线	决定关键帧之间"怎么走"的曲线与它在关键帧处的陡峭程度	5.4
	步进（阶跃）	保持不动、到下一关键帧瞬间跳变——Blocking 阶段的标准插值	5.4
	摄影表（Dope Sheet）	只看关键帧"落在第几帧"的视图，用于整体调整时间与间距	5.4
	Blocking / Splining / Polishing	关键姿势布局 → 顺曲线 → 打磨这三段标准流程	5.4
	循环与往返（cycle / ping-pong）	循环动作的出入范围类型：首尾相接循环，或往返摆动	5.4
	动画十二原则	预备动作、跟随与重叠、缓入缓出、弧线、时间节奏等十二条运动观感法则	5.6
IK / FK 与重定向	**FK 正向运动学**	转关节、运动沿链条逐级传递：输入角度，输出位置	5.5
	IK 反向运动学	把链条末端拖到目标、反解各关节角度：输入位置，输出角度	5.5
	烘焙动画（Bake Animation）	把约束与 IK 的结算结果固化成逐帧的 FK 数据，导出前必做	5.5
	运动条件	动作数据能被顺利套用的前提：命名、比例、帧率、根运动约定	5.7
	动作重定向（motion retargeting）	把动作从源骨架"翻译"到目标骨架：旋转可搬，位置要换算	5.7
材质与修复	**UV 与间距 / 纹素密度**	表面到二维的展开图；岛间要留空隙，像素与面积之比要全模型一致	5.9
	法线贴图	记录表面朝向微变化的贴图；注意方向约定，禁止 AI 去噪	5.9
	PBR	基于物理的材质体系：底色、法线、粗糙度、金属度等分通道描述	5.9
	色彩空间	颜色类贴图按 sRGB、数据类贴图按 Non-Color 导入	5.9
	重拓扑	把生成网格整理成面数合理、形变区布线合格的可动网格	5.9
	修复优先级	几何 → UV → 法线 → 贴图，从根修到叶，顺序不可反	5.9

5.11 动手实践：「举杯邀月」从绑定到交付

本周三个学时，两个用于实践。目标明确：**让第 4 章做好的诗人角色，完成一个 4–6 秒的「举杯邀月」动画片段，并留下一份可以自证清白的记录。**春江小组的分工是这样的：小舟负责模型导入、体检与绑定；小月负责动作获取与套用；两人一起做人工修正——理由很朴素，修正这段最需要"两个人的眼睛"，因为自己盯久了看不出问题。

先看这段动画的"设计图"：关键姿势怎么摆。任何时候你都可以回到这张图来对照——它也是你人工修正阶段的"参考答案"。

图 5-14　「举杯邀月」关键姿势时间线。这张图就是你的"参考答案"：当套用来的动作让你觉得"哪里不对"时，回到这里逐个姿势对照——缺的是哪个关键姿势？错的是哪段节奏？

下面是全程的八步工作流。注意看每一步标出的"谁来做"：自动化的环节越来越少，需要判断的环节越来越多——这就是 2026 年角色动画工作流的真实形状。

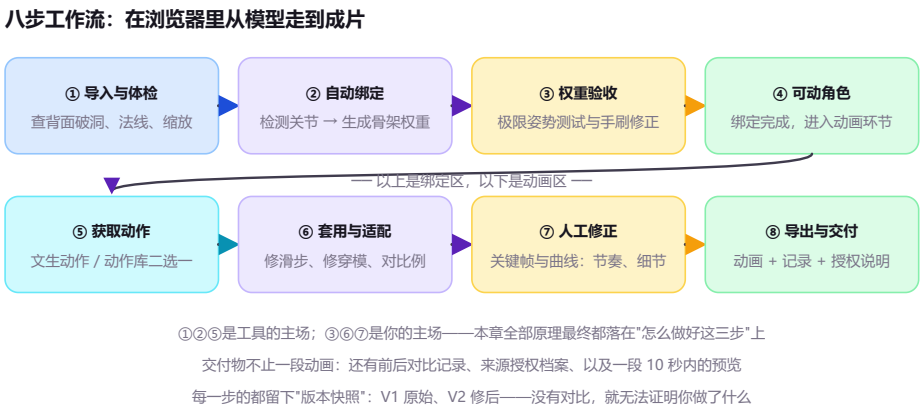


图 5-15　八步工作流。闭环里标色的三个环节（权重验收、套用适配、人工修正）是本章考核的真正重点：它们不可能被"点一下按钮"完成，也正是这门课为什么值得学的理由。

图 5-16　主线角色：诗人（设定图）。这是第 4 章资产进入本章前的「标准像」：站姿精神、服装结构清楚、无遮挡。提醒：设定图好看≠能绑定，导入后的体检（轴向、尺度、原点、法线）一步都不能省。 AI 生成

🛠 实践一：导入体检与自动绑定（40 分钟 • 课堂）

1. **导出与导入。**从第 4 章的成果里导出诗人角色的 GLB（或从你的资产库选取），导入网页绑定工具。导入后先做"一分钟三查"：**缩放**（身高是否合理量级，别是厘米/米混用）、**原点**（脚是否站在地面上、坐标原点在脚下而不是身体中央）、**法线朝向**（开启背面剔除转一圈，看有没有"黑洞"）。
2. **姿势准备。**确认角色是 T-pose 或 A-pose；若手臂贴着身体，先摆开——自动绑定的关节检测对遮挡很敏感。
3. **执行自动绑定。**选择人形绑定（若工具支持"检测置信度"，先看一遍；置信度低的关节往往是出错的高发区）。记录：工具名、耗时、输出格式。
4. **权重验收。**按表 5-3 的清单逐项测试——把肘、肩、膝、髋各自弯到极限姿势，观察塌陷与渗漏；重点看腋下、裆部、衣摆。
5. **手刷修正（尽力而为）。**对发现的问题区域用权重画笔修正；修不完的写进问题清单——它会成为第七步的重点。

基础

完成绑定并导出，得到"带权重可动"的文件。截图记录：绑定前的体检结果 + 极限姿势测试的四个姿势各一张。

进阶

用同一角色尝试**两个**工具的自动绑定（如一个网页工具 + 本地免费工具），对同一组极限姿势截图对比：谁的腋下更干净？谁的衣摆更可控？把对比写进创作档案。

🛠 实践二：获取动作并套用到角色（40 分钟 • 课堂）

1. **路线甲——文生动作。**按提示词卡（5.8.2）的方法，用"结构化英文 + 精确时长"的思路生成「举杯邀月」动作；只写动作、不写修饰；时长控制在 4–5 秒。
2. **路线乙——动作库。**在 Mixamo 里搜索举手、饮酒、致以敬意的片段，挑选 1–2 段；下载时注意选择与你的骨架匹配的标准格式。
3. **套用与重定向。**把动作应用到诗人的骨架上。观察四个"运动条件"哪个没满足：命名有没有自动对上？比例差异大不大？时长与原始设计相符吗？位移（根运动）符不符合你要的"原地"？
4. **首轮质检。**录屏回放三遍：脚有没有滑（侧视看支撑脚）、手有没有穿模（近景看胸口）、整体节奏对不对（对照图 5-14）。
5. **版本快照。**把"原始套用版"命名为 V1 存盘——无论多糟都存，它是你对比记录的起点。

基础

得到一段能播的「举杯邀月」V1；写三条"最明显的问题"（越具体越好：发生在第几秒、哪个部位）。

挑战

同一条描述分别用"文生动作"和"动作库"两条路线各做一版，对比：哪一版的"表演感"更好？哪一版更容易修？写 100 字以内结论。

🛠 实践三：人工修正——把"能动"变成"会演戏"（40 分钟 • 课堂 + 课后）

这一步是本周的核心。修正的基本动作有两个：**改姿势**（重摆关键帧）与**改节奏**（在曲线视图里调整缓入缓出、加减停顿）。按下面的顺序做：

1. **修接触问题（技术优先）。**先解决滑步与穿模——它们不修掉，后面的美感讨论没有意义。支撑相里脚必须"钉"住不动；手臂贴胸的动作留出安全缝隙。
2. **补预备动作。**在举起之前插入"沉肩蓄力"的关键姿势（图 5-14 的 ②）：身体先微微下压后仰，再启动。对比赛：加与不加，各导出 3 秒预览，放给同伴看——让他指出哪一版"更想跟着使劲"。
3. **调节奏。**把"举起"段改为两头缓、中间快；在"仰头望月"处加一个明显停顿（比如 0.5–0.8 秒几乎不动），让观众有时间读"望月"这个动作。
4. **修弧线。**打开运动轨迹或逐帧看手的路径：是弧线还是折线？折线处补中间帧或调整切线。
5. **加跟随与重叠。**给衣袖/飘带留"欠半拍"的跟随：主体的手停住后，衣袖再经过 2–4 帧落停。没有物理工具的，就手动在衣摆骨架上偏移几个关键帧。
6. **导出 V2 并做对比评估。**填写表 5-15 的前后对比记录；导出最终预览（10 秒内）。

修正前 • V1 的典型症状	修正后 • V2 的目标状态
举杯像"被吊上去的"：匀速上升、没有蓄力，手到胸口才突然减速。	蓄力 → 举起 → 停顿 → 回落，四段清晰；举起的曲线两头缓、中间快。
仰头与举杯同时结束——没有停顿，观众来不及读"望月"。	"仰头望月"处停住一拍；手走的是弧线而不是折线。
手停下之后，衣袖"啪"地一起静止，像一块纸板。	手停后衣袖再"欠"2–4 帧落定，整段动作有呼吸感。
侧视可见：支撑脚在地面上缓缓前移（滑步）。	支撑脚全程钉住；手腕与胸口保持安全距离，无穿模。

🛠 实践四：交付与归档（20 分钟 • 课后）

交付物不是"一个文件"，而是一份能自证的"作品包"：

- ▣ 交付 1 绑定完成的角色文件（含骨架与权重）——命名规范、注明版本。
- ▣ 交付 2 「举杯邀月」动画文件 + 10 秒以内的预览视频（含 V1 与 V2 两版，标注哪版是终版）。
- ▣ 交付 3 表 5-15 的前后对比记录表（填写完整，含五维评分）。
- ▣ 交付 4 来源与授权档案：模型的生成工具、动作的来源（文生动作的参数 / 动作库链接与许可）、绑定工具与设置。
- ▣ 交付 5 AI 参与声明与标识：哪些环节由 AI 完成、哪些由人修正；对外发布时按第 1 章的要求添加生成标识。

图 5-17　「举杯邀月」的表演参考。把这张图读成三件事：①姿态——重心在一侧腿、上身微仰、举杯的手腕有一道自然的翻折；②节奏——袖子甩起的弧线说明「到这里应该是慢的」；③预备动作藏在哪里——没有它，举起的手会像被吊上去的。 AI 生成

图 5-18　角色动作的时间轴（视频）。把这段 5 秒的动作用本章的「修正口诀」过一遍：**预备动作**有没有（举杯前身体先有轻微的下沉）、**跟随与重叠**有没有（衣袖比手臂晚半拍停下）、**缓入缓出**有没有（起手与落定处是否顺）。对着视频逐条找问题，比对着单帧静图找问题有效得多——这正是动画行业的审片方式。 网页版可播放 AI 生成

◆ 修正口诀：一剪、二停、三缓、四跟、五查

人工修正听起来复杂，但在本章任务的尺度上，它可以压缩成五个字的操作口诀——按顺序执行，每一字对应一个具体的动作，做完就过一字，又快又不容易漏：

一剪：先做减法。把明显多余的关键帧删掉——AI 生成的动作往往"关键帧过密"，导致节奏琐碎、动作发飘；减少到五到七个关键姿势，动作立刻变干净。这一步和 5.4 节说的"先定姿势"是同一个道理：不是越多的关键帧越精细，而是越少的关键帧越需要你思考。

二停：在任何你想让观众"读一下"的地方加停顿——仰头望月的瞬间、杯到胸口的一拍。停顿时长不必长，零点几秒的"几乎不动"就能让动作产生"呼吸"。这是回报率最高的一笔修改：加停顿的成本是一两分钟，观感的提升立竿见影。

三缓：把匀速段改成缓入缓出——如果工具不给曲线控制，就用刚才说的"加中间关键帧"办法。检查标准是 5.4 节那句话：真实世界里几乎没有东西是匀速启动的。

四跟：给附属物（衣袖、飘带、头发）设置"晚停 2–4 帧"的跟随——主体停下之后，它们还在"欠"着走。没有物理模拟就手动偏移关键帧，效果一样成立。

五查：最后回到表 5-16 的五维自查，逐维过一遍；把每项评分与理由写进表 5-15 的对比记录。口诀的顺序也不能乱：先剪干净再调节奏，先调主体再管附属——和"修复优先级"一样，从根到叶。

表 5-15　动画质检五维：每个版本交付前自查一遍

维度	怎么检查	合格线	常见扣分点
① 姿态与意图	定格看每个关键姿势的截图	不看前后文也读得懂"这是举杯邀月"	姿势含糊、手脚位置别扭、剪影不清楚
② 接地与接触	侧视看支撑脚的落点是否稳定；看手与道具的接触	支撑相内脚不滑动；手与杯子不"脱手"、不穿模	滑步、脚陷地、手腕穿胸口
③ 节奏与曲线	看曲线视图；数一数每段占几帧	没有整段匀速；发力处干脆、收势柔和	全程匀速、该停的地方没停、抖帧
④ 细节与跟随	慢放看衣袖、头发、飘带的停止时间	附属物晚于主体停下 2–4 帧；弧线正确	衣物僵硬、与身体同步静止、轨迹折线
⑤ 整体可信度	关掉声音正常速度播放三遍，问同伴"像不像人"	同伴的第一反应是"顺"，而不是"哪里怪怪的"	"橡皮人"、无预备、重量感缺失

表 5-16　「举杯邀月」前后对比记录表（模板，复制进你的创作档案）

版本	来源与设置	人工改了什么	五维评分（各 1–5 分）	结论
V1 原始套用	（工具 · 动作来源 · 参数）	——	–	作对比基准
V2 修权重/接地后	（绑定工具 · 权重修正要点）	修了渗漏；脚重新对齐	–	保留
V3 关键帧修正后	（曲线调整 · 关键帧增删）	加预备动作；加停顿；修弧线	–	候选终版
终版	（导出设置 · 文件格式）	衣摆跟随；整体打磨	–	**交付**

5.12 课程思政：AI 辅助，不等于放弃艺术标准

本章请你记住的技术结论有很多，但这一节，只讲一件事——它比任何技术都重要：**当工具把"从 0 到 1"的成本压到近乎为零时，决定作品命运的，是你对"好"的坚持。**

◆ 5.12.1 行业正在发生的变化：从业者的位置在移动

本章记录的这些工具——自动绑定把权重计算压到约 30 秒、文生动作把一段表演的"第一稿"变成一句提示词、动作库把真人动捕免费开放——不是未来的想象，是此刻正在发生的变化。看动画行业里从业者的公开讨论，能发现一个耐人寻味的现象：被从业者反复点名"希望 AI 尽快接管"的，恰恰是那些"无聊的部分"——干净的 UV、正确的法线烘焙、自动材质分配、骨骼权重；而"有意思的部分"——表演怎么设计、节奏怎么打、哪一帧该停——依然被牢牢握在动画师手里。

这给我们的启示不是"哪些岗位要消失"，而是"**人的价值在向哪里迁移**"：从"把活做出来"迁移到"判断活做得对不对、好不好"。本章的技术训练，正是为这个迁移做准备——权重验收、滑步诊断、曲线修正，这些听起来琐碎的动作，恰恰是"判断力"最具体的形态。

◆ 5.12.2 工匠精神：把"返工"当成作品的组成部分

本周的实践环节会让每个人经历一件"不体面但重要"的事：AI 给你的动作第一版几乎一定不能直接用——脚在滑、手在穿、节奏是平的。此时有两条路：一条是把 V1 直接交上去，在记录里写"AI 的效果就这样"；另一条是走完本章的修正流程，把 V1 修成 V2。

选择第二条路的人，会在某个时刻体验到一个微妙的瞬间：**当你为了"半拍的停顿"反复拖拽曲线、当衣袖终于"欠半拍"地停住时，你会突然明白这段动画里哪些部分是"你的"。**那种手感、那份判断、那版对比记录，就是工匠精神的现代版本——它不神秘，就是"愿意为看不见的细节花时间"。审美判断必须花时间，而时间就是尊重的形式：尊重作品、尊重观众，也尊重你自己的名字。

◆ 5.12.3 尊重劳动与守住边界：素材背后的三个提醒

AI 让获取素材变得廉价，但素材背后的劳动与权利并没有消失。三条提醒，请你带进每一次创作：

- **动作数据是别人的劳动。**动捕库里的每一段数据都来自真人表演者与动捕团队，许可条款（如 CMU 的"可商用、不可转售、需致谢"）就写在这些劳动的边界上。用之前读许可原文，用之后按规范致谢——这不麻烦，三两分钟的事。
- **生成内容的署名与标识是义务。**用 AI 生成的模型、动作与贴图，对外发布时按课程与法规要求添加生成标识、如实说明 AI 参与程度。把这条当成作品的"格式条款"，而不是"额外负担"。
- **别把"快"当成唯一标准。**本章的每一项技术都在让你更快——更快地得到动作、更快地修完缺陷。但请始终记得：快，是为了把省下来的时间投入到判断里去；如果省下来的时间什么也没换来，那我们就只是"更快地生产了平庸"。

🇨🇳 思政一　创作者劳动与审美判断：谁为"好"负责

为什么必须讲这个。逐帧看一段 AI 生成的动作时，你会经历本章反复描述的那种体验："动作是对的，但它不是好表演。"预备动作缺席、停顿时机不对、衣袖像纸板——每一项都不是"技术故障"，而是**审美判断缺席**留下的空白。工具的职责是生成"运动"，审美的职责是生成"表演"，二者之间有本质区别。AI 降低了"制作"的门槛，却把"判断"的门槛留在了原地——甚至抬高了，因为选择变多了。

在本课程里，这意味着什么。你的作业评分表里有一条硬性要求：**提交 V1 与 V2 两个版本，并写清你在两版之间做了什么、为什么这样做。**这条要求考核的不是"修得好不好"，而是"有没有判断、能不能说出判断的依据"。哪怕某处改动最后被证明是错的，只要你能说清"我以为会更好，因为……"——那也是一次有价值的创作决策。反之，把 V1 原样交出、写一句"AI 就这样"，无论画面多流畅，都不算完成创作。请和你未来的职业身份对一下表：动画师、设计师、导演，签在作品上的从来不是工具的名字，而是你的名字。

写给你的长期提醒。工具会一直变强，"让 AI 全自动完成"的诱惑会一直变大。但只要你在这个行业里做一个"署名的人"，就请守住这条底线：**可以放手给工具的，是生产；不能放手给工具的，是判断。**本章的失败诊所里，每一个"橡皮人""滑步的诗人""背上的窗户"，都不是工具的错——是没有人替"好"站岗。

🇨🇳 思政二　工匠精神：从"能用"到"传得下去"

为什么必须讲这个。你可能会问：一个课程作业而已，为什么要花两个小时修"半拍的停顿"和"衣袖的延迟"？答案藏在动画这门手艺的历史里。从迪士尼的九位老画师到今天的三维动画师，"一帧一帧抠出来"的传统从未断过——正是这种对细节的"不放过"，让那些作品几十年后依然活着。**能快速做出来的东西，往往也快速被忘记。**我们这个时代最稀缺的，不是生成能力，而是"愿意在生成之后继续坐下"的耐心。

在本课程里，这意味着什么。把工匠精神拆成三条可以在本周执行的动作：① **做完再离开**——导出前跑完表 5-16 的五维自查，不带着已知问题交付；② **留痕**——保留 V1 与 V2 的对比记录，让"我贡献了什么"有物证；③ **修到"下一个问题不再值得修"为止**——工匠精神不等于无限打磨，而是知道"这一版的瓶颈在哪里、值不值得继续"的分寸感。这种分寸感，正是职业与业余的分界线。

写给你的长期提醒。你未来会大量使用 AI 生成资产与动作——这不丢人，这恰恰是这一代创作者应有的工作方式。丢人的只有一种情况：**作品上有你的名字，而作品里没有你的判断。**愿你的每一次"导出"，都经得起自己回看第三遍。

◆ 把这一章压缩成七句话

1. 动画的本质：**姿势 × 时间。**3D 动画师摆关键姿势、让计算机插值；"动画在动什么"这一问，答案永远是"关节角度随时间的集合"。
2. 骨骼是一棵层级树：**运动从根向叶传递**——转肩带动整条手臂，转肘只带动前臂与手；附肢的层级与自由度，决定了角色"能做什么、不能做什么"。
3. 蒙皮是把顶点绑定到骨骼，**权重定义每个顶点"听谁的"**；权重渗漏是自动绑定后最常见的质量问题——"橡皮人"、抬手拽全身，根源都在这里。
4. 关键帧管"到过哪些姿势"，**曲线管"怎么过去"**；匀速显得机械，缓入缓出才有重量感；制作顺序是 Blocking → Splining → Polishing，顺序不可反。
5. FK 转关节（角度输入、位置输出），IK 拖末端（位置输入、角度反解）；IK 可能无解或多解，导出前要烘焙成 FK，且"不要过度使用 IK"。
6. 动作重定向像翻译：**旋转可以直接搬，位置必须按比例换算**；滑步的根源是几何（腿长不同、同样的角度送不到同一地点），不是软件"没算对"。
7. AI 工具的现状是三块拼图——自动绑定、文生动作、动作库；它们把"从 0 到 1"压到分钟级，但"从 1 到能交付"（权重验收、适配修正、关键帧打磨）仍然是人的工作；材质修复同理：从几何修到贴图，顺序不可反。

◆ 自测

❓ 自测 1：用"姿势 × 时间"解释：为什么三维动画里"中间帧"不再需要人来画？这给创作者带来了什么好处、什么新的责任？

答案：动画可被分解为"关节角度随时间的函数"。三维里关节角度是数值，关键帧给出若干时间点上的数值，中间帧就成一个纯粹的数学问题——按插值曲线补出连续值，计算机瞬间完成。好处是"体力活"被自动化，创作者可以把精力集中在关键姿势的表演设计上；新的责任恰恰在同一个地方：**中间怎么走（曲线）变成了你要负责的创作决定**——匀速还是缓入缓出、在哪儿加停顿，这些决定动作的"重量感与生命感"。工具替你算了，但没替你判断。

❓ 自测 2：一位同学做完自动绑定后说"角色能动能摆姿势，没问题"。请用"加权平均"解释：为什么"能动"远不等于"绑定合格"？请给出至少两个必须检查的位置。

答案：蒙皮的每次形变都是"每根骨骼提出一个顶点该去哪的方案，再按权重加权平均"。能摆姿势只说明"平均"这件事在算，不说明权重分配是对的：一个顶点若被错误地分到远处骨骼的权重（渗漏），平常姿势下看不出来，一旦关节到极限位置，错误权重就会把它拽走——橡皮人、抬手拽全身都是这么来的。必须检查的位置包括：腋下（手臂与躯干几何靠近）、裆部（两腿之间）、肘与膝（过渡带宽度）、衣摆与飘带（附属部件）、手指与面部（通常需要专门绑定）。检查方法不是"看颜色是否均匀"，而是把关节弯到极限位置观察。

❓ 自测 3：两段动画的关键帧完全相同（同样的姿势、同样的时间点），观感却完全不同——一段像机器人、一段像真人。用本章的概念解释这是怎么回事，并说出一种生活里的同类现象。

答案：区别在插值曲线。关键帧只规定"经过哪些姿势与时刻"，"怎么经过"由曲线（切线）决定：全程线性（匀速）→ 机械感；两端慢中间快（缓入缓出）→ 自然、有重量。生活里的同类现象：电梯的启动与停止都有缓冲，若匀速直上直下，乘客会觉得像被机械臂拎着——因为现实中没有质量的物体才匀速，克服惯性必然带来"渐入渐出"。这也解释了行业说法"匀速运动会显得机械，曲线视图是核心工具"。

❓ 自测 4：下面两个任务，你会选 FK 还是 IK，为什么？① 挥手告别的整段动作；② 把手掌精确按在琴键上，同时身体前倾。各说一条理由，并说出 IK 的两个"脾气"。

答案：① 选 FK——任务的"味道"在关节的角度节奏上（手臂摆动的弧度与快慢），逐关节调节奏最直接。② 选 IK——任务要求末端精确对位（按在琴键上），把末端"钉"住再动身体，省掉大量试错。IK 的两个脾气：**无解**（目标超出可达范围，两边之和 L_1+L_2 小于距离 d 时，三角形不成立，只能伸直或拉伸）与**多解**（三角形有两个解，肘朝上或朝下，需要极向量告诉它方向）。此外导出前要把 IK 烘焙成 FK，否则下游引擎无法重现。

❓ 自测 5：把一段"标准成人比例"的走路动作套到腿更短的角色上，出现滑步。请用几何解释原因，并说出两条处置办法。为什么"旋转可搬、位置要换算"？

答案：原因：动作数据里的骨盆高度与步幅是按源角色腿长组织的，腿短的角色沿用同样的角度，脚到不了"应该在的落点"——几何上，骨骼长度变化后，同样的关节角度把末端送到不同的世界坐标。处置：① 按腿长比例缩放位移与步幅（重定标落点）；② 用脚锁定把支撑脚"钉"住，重排落地关键帧；必要时换一段动作或自 K 关键帧。为什么旋转可搬：旋转描述"朝向的变化"，与身高无关，只要骨架拓扑对应，角度数值可以直接平移；位置有量纲、依赖骨骼长度与骨盆高度，必须按比例换算。

❓ 自测 6：拿到一个 AI 生成的角色资产，正面贴图精美、背面有破洞、UV 拉伸、AI 生成的金属度通道发灰。按"修复优先级"排出你的工序，并解释为什么顺序不能反。

答案：工序：① 几何层——先补破洞、清理内部面与翻转法线（背面破洞不修，后面全部白做）；② UV 层——重展开、统一纹素密度、留足间距与烘焙 margin（UV 一变，贴图坐标全变）；③ 法线层——重建高低模关系后重烘法线，核对方向约定，不做 AI 去噪；④ 贴图层——最后才重绘画接缝、修光照、手工重画金属度（建议接近二值的黑白分区）与粗糙度。顺序不能反的原因：几何是根、UV 是干、贴图是叶——在破网格上重画贴图等于"把画贴在漏风的墙上"；UV 改动会让所有贴图工作作废。一句话记法：从根往叶修，前一层没修完，不要碰后一层。

◆ 拓展阅读

- **随查手册：**本章涉及的工具（自动绑定、动作库、材质修复）选型与额度见附录 B 工具速查；克隆真人动作/形象时的授权与标识要求，见附录 E 合规清单与授权书模板。
- Ollie Johnston 与 Frank Thomas《生命的幻象》（The Illusion of Life: Disney Animation）——动画十二原则的原典；建议先读"预备动作""跟随与重叠""缓入缓出"三节，再回头重看本章 5.6。
- Richard Williams《动画师生存手册》（The Animator's Survival Kit）——节奏与表演的经典教材；虽是手绘体系，但"关键帧 + 中间帧"的思维方式与三维完全相通。
- Blender 官方手册的骨骼（Armature）与蒙皮章节 + Rigify 插件文档——免费、系统；即使你主要用网页工具，读一遍也会让"权重、约束、烘焙"这些词真正落地。
- CMU MoCap 数据库（mocap.cs.cmu.edu）与 Mixamo 帮助中心的许可说明——请读一次原文，练习"读许可"这个职业习惯；你会发现 90% 的版权问题都源于"没读条款"。
- HY-Motion 1.0 的 GitHub 仓库与技术报告（arXiv 2512.23464）——想深入了解文生动作的读者，从"数据管线"一节读起：它解释了为什么模型"不支持情绪与服饰"这类输入。
- 3ds Max 或 Maya 官方动画教程中关于 IK/FK 混合、烘焙动画与摄影表的章节——本章 5.4、5.5 的概念在这些教程里有成套的操作演示，可作为课堂之外的延伸练习。

第 6 章

人体姿态与动作迁移

让真人跳一遍，角色就会了——关键点、置信度、平滑与重定向的完整链路。

学习目标

1. **知识：**理解二维姿态估计的关键点定义（COCO 17 点、MediaPipe 33 点等）、热图与置信度的含义、三维姿态与 SMPL 类人体模型的基本概念，以及遮挡与抖动的成因与处理思路。
2. **能力：**能拍摄规范的参考视频并提取姿态数据；能用"动作参考 / 姿态条件"把真人动作迁移到生成的图像角色上；能把动作重定向到第 5 章做出的三维角色上；能用一份排查清单定位滑步、穿模、比例失真等问题。
3. **素质：**形成"身体影像先授权、再采集、后标识"的工作习惯；理解深度合成的新规与"可识别性"标准；尊重不同民族、不同群体的身体文化表达方式。

对应第 7–8 周 · 6 学时　讲授 3 学时 · 实践 3 学时　主线推进：捣衣 / 拂袖动作迁移

> 周五下午的舞蹈教室，小舟把手机架在三脚架上，请小月示范了一遍《春江花月夜》里最难处理的那个动作：捣衣——抬臂、落袖、沉肩、再回手。三十秒的素材，约 900 帧。三小时后，这段真人的动作出现在了两个地方：一处是图像工具生成的"捣衣仕女"图，人物的手臂与袖口与真人几乎同步；另一处是第 5 章做的诗人角色，他也在做同样的动作——只是脚下有点打滑，右手袖子穿过了身体。

从"真人动了一下"到"角色动了起来"，中间发生的事比大多数人想象的多。模型并不是在"看视频"，它是在**读一串坐标**：把每一帧里的人体压缩成十几个点，给每个点一个"我有多确定"的分数，把这些点连成骨架，再想办法把这套骨架搬到另一个身体上。这一章要回答三个问题：**人的动作是怎么变成数据的？为什么这些数据会抖、会错、会丢？以及怎么让另一个"身体"不失真地把它演一遍？**前者是姿态估计（pose estimation），后者是动作迁移（motion transfer）与动作重定向（motion retargeting）。

📌 本章的四族技术：别把它们混成一件事

本章的对象是「人的动作数据」，但围绕它站着**四族完全不同的技术**——输入输出、失败模式、验收标准各不相同：

- ① **姿态估计**（6.1–6.3）= **判别**：从视频里「读出」关键点坐标——解决「数据从哪来」。看置信度与遮挡。
- ② **姿态条件生成**（6.4）= **生成**：用骨架图控制图像生成——解决「动作怎么变成画面」。看姿态还原度。
- ③ **动作重定向**（6.5）= **几何换算**：把动作「翻译」到另一副骨架——解决「动作怎么交给三维角色」。看滑步与穿模。
- ④ **参考动作驱动**（6.6）= **生成**：一条参考视频直接驱动角色——解决「一条视频、一次模仿」。看身份保持。

本章开头的图 6-1 说的是②与③「两条去向」的区别；①②③④ 合起来，才是这份数据从拍摄到成片的完整去向。出了问题时，先定位是哪一族出的错，再谈怎么修——这正是本章反复练的排查习惯。

6.1 姿态：把一个人的动作变成一串坐标

先请你做一个小实验。打开手机相册里任意一段有人物走动的视频，盯着看十秒，然后问自己：**你是怎么"看懂"这个人在走路的？**你大概会答"因为他的腿在交替摆动"。但请注意，你从来没见过"腿在交替摆动"这个东西本身——你看到的是像素颜色在变化，是"腿"这个概念从你的经验里被自动调用了出来。计算机没有这种经验，所以它需要一个中间层：**把人体抽象成十几个关键点（keypoint），把动作抽象成这些点随时间移动的轨迹。**这就是本章所有内容的起点。

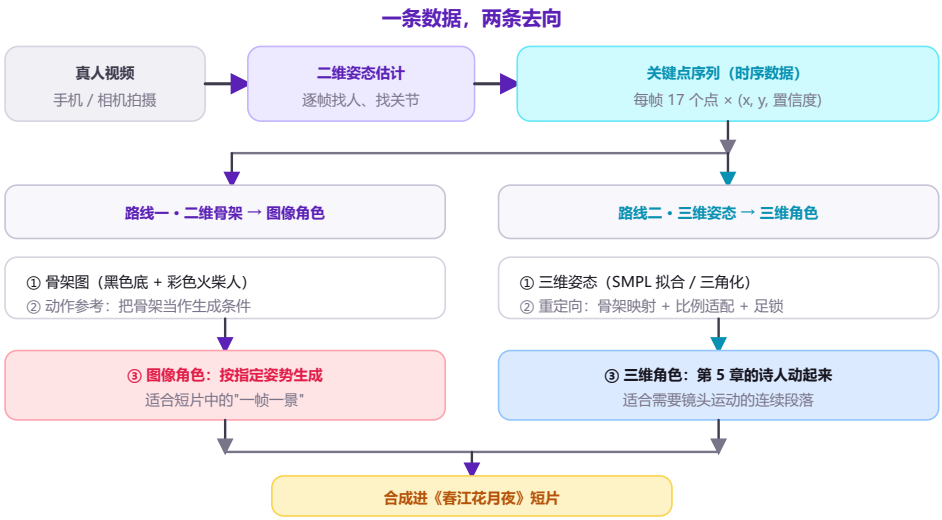


图 6-1　本章技术路线图。同一份关键点数据，向左走是"图"（姿态条件生成静止画面），向右走是"骨架"（重定向驱动三维角色）。两条路线的差别不在数据来源，而在"谁来解释这份数据"。

关键点：人体的"降维打击"

姿态估计（pose estimation）指的是从图像或视频中定位人体关节位置的任务。它的产物不是一张图，而是一组**关键点**（keypoint）：每个点用一对坐标 (x, y) 表示，外加一个表示"模型有多确定"的分数。把一个人压缩成十几个点，信息量损失了 99% 以上——但剩下的 1% 恰好是"动作"本身。衣服的颜色、脸的长相、背景的陈设全被丢掉了，留下的是姿势与运动的抽象。

这个思路在生活中其实很常见。健身镜上跳动的火柴人、体感游戏里跟着你动的小人、运动 App 里的跑姿分析曲线——它们背后都是同一件事：先把人变成点，再对点做计算。对创作而言，这个"降维"带来两个直接好处：**第一，它让动作可以脱离具体的人被保存和搬运**（不涉及长相与身材，搬运起来没有负担）；**第二，它给了生成模型一个可以精确控制的把手**——你想让画面里的人抬手 45 度，与其用文字描述半天，不如直接在骨架图上把那个点拖上去。

那么，一个"人"到底该用多少个点来表示？这不是可以拍脑袋决定的问题：点太少，动作表达不清（只有肩和腕两个点，就分不清手肘是弯是直）；点太多，标注成本暴涨、模型也更容易出错。业界因此形成了若干套"约定"。本书需要你记住其中四套——不是因为要背数字，而是因为你在不同工具之间搬运数据时，一定会遇到"我这边是 33 点，你那边是 17 点"的转换问题。

表 6-1　四套常见的人体关键点体系

体系	点数	构成	典型使用场景与特点
COCO	17	鼻、双眼、双耳、双肩、双肘、双腕、双髋、双膝、双踝	学术基准与工程事实标准，被 HRNet、YOLO-pose 等广泛采用；只覆盖身体主干，不含手指与面部细节
MediaPipe Pose（BlazePose）	33	身体点细分，另含手部与脚部相关点	面向移动端与实时场景，算力需求低；每帧输出归一化的 (x, y, z) 外加一个可见性 / 置信度字段
DWPose（COCO-WholeBody）	133	17 身体 + 68 面部 + 42 手部（每手 21）+ 6 脚部	全身检测精度高，全身 AP 约 66.5%；手部表现明显优于早期方案；Apache-2.0 许可，可自由商用
OpenPose	135	身体 / 脚 / 脸 / 手	开山之作，单阶段自底向上，靠"部位亲和场"连接关节点，无需单独的人体检测步骤；但技术较老、商用许可受限

四套体系的关系可以用一句话概括：**COCO 的 17 点是"普通话"，其他体系是带方言口音的变体**。COCO 17 点只覆盖身体主干，因为它的设计目标是"一根能表达动作的骨架"；MediaPipe 的 33 点在同一根主干上增加了面部与手部的粗略点，因为它的目标场景是"手机前置摄像头里的实时互动"；DWPose 与 OpenPose 的 133 / 135 点则把手指的每一个关节、面部的每一个轮廓点都标了出来，因为医疗、驱动虚拟主播等高精度场景需要它们。

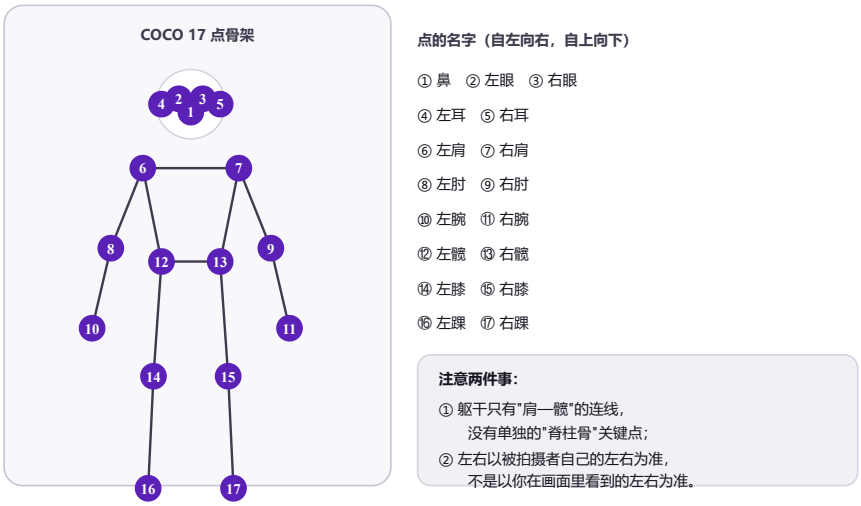


图 6-2　COCO 17 点骨架。这是本章出现频率最高的一张图：后面讲热图、置信度、平滑、重定向，对象都是这 17 个点。请特别注意右下的两句话——它们分别对应 6.1 节的映射问题与 6.7 节的"镜面翻转"失败。

图上那两行小字值得展开。**第一，"没有脊柱骨"**：COCO 只要求"肩"和"髋"两组点，从肩到髋的连线是隐含的，模型并不输出一个"腰椎"点。而多数虚拟角色的骨架是含脊柱的（颈、胸、腰若干节），所以把 17 点直接贴到虚拟角色上时，必须自己"补"出脊柱的运动——这是 6.2 节要解决的问题。**第二，"左右以被拍摄者自己的左右为准"**：如果画面里的人背对着你，他抬起的右手出现在画面左侧。"左"永远指被拍摄者自己的左边；一旦搞混，迁移出来的动作就会左右互换，这正是"镜面翻转"这类失败的温床。

◆ 热图与置信度：模型真正输出的是什么

很多教程会告诉你"姿态估计就是输出 17 个坐标"，这句话在工程上不够准确。绝大多数模型并不直接吐出坐标，而是先为**每一个关键点生成一张"概率地图"**——**热图**（heatmap）。热图是一张与原图（或降采样后的特征图）同尺寸的灰度图，图上每个位置的数值表示"这个关键点出现在这里的可能性"，最高点就是模型认为的关节位置。

用一个类比：想象班里要选出"最像班长的同学"，老师没有直接点名，而是给每个同学打了一个"像不像"的分。热图就是这张打分表，而最终坐标只是"分数最高的那位"。这样设计有两个好处：**第一，模型学起来更容易**——预测一张平滑的打分表比一步到位预测一个精确坐标要稳定得多，因为"附近也很像"这件事在打分表里可以被表达出来，而坐标只能表达一个点；**第二，热图的形状本身就携带了不确定性信息**——一个尖而高的峰意味着"非常确定"，一个矮而扁的峰意味着"大概在这一带，但我拿不准"。

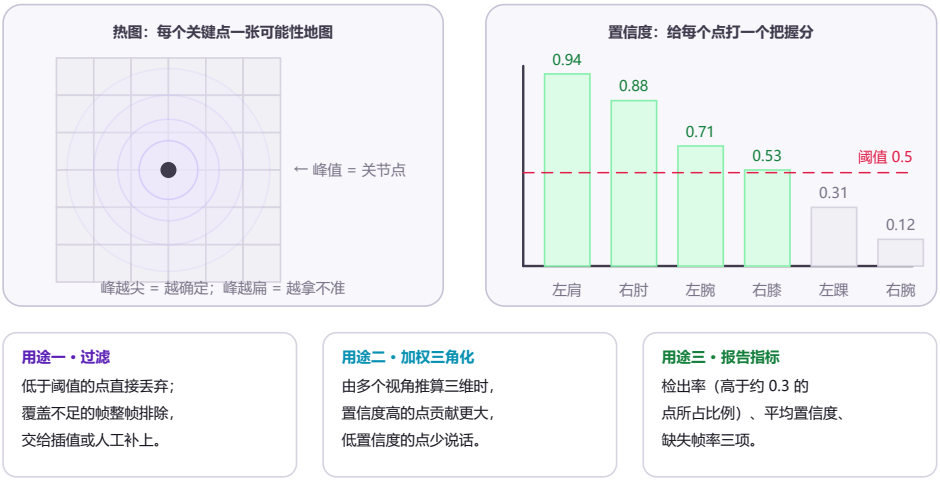


图 6-3　热图与置信度的两种视角。左图看"位置从哪里来"，右图看"这个位置可不可信"。注意右图 0.53 与 0.5 只差一点——**阈值附近的点最危险**，它们会在相邻帧之间反复跨越阈值，表现为"点时有时无"的闪烁。

图中第三个数——**置信度**（confidence）——是本章最容易被忽略、却最影响成品质量的一个量。它的确切含义**因模型架构而异**：可能来自热图的峰值高度，可能是另一个分支网络算出的分数，也可能是像 MediaPipe 那样直接给出一个"可见性"（visibility）字段。你不需要知道它内部怎么算，但必须知道它被用来做什么，否则排查问题时会对着一个"明明看得见却画错位置的点"束手无策。

表 6-2　置信度的三种用途与一条警告

用途 / 读法	具体做法	对你作品的直接意义
过滤	置信度低于约 0.5 的关键点直接丢弃；关键点覆盖不足的帧被排除	决定了"这一帧还能不能用"。丢弃之后要补插值，**补得好不好直接决定动作是否连贯**
加权三角化	由多个二维视角推算三维时，用置信度作为权重，高置信度点对三维解的贡献更大	单目 / 多目视频做三维姿态时的核心技巧；权重给错，深度就会"漂"
报告指标	检出率（高于约 0.3 的关键点比例）、平均置信度、缺失帧率	拍摄完成后先看这三个数，再决定要不要补拍——**比肉眼逐帧检查快得多**
警告 · 它不是准确率	置信度只回答"模型自己有多确定"，不回答"它对不对"	遮挡场景里模型可能"非常自信地把手画在腰上"——**高置信度的错误比低置信度的错误更危险**

> ⚠️ **常见误解**
>
> 很多同学第一次看到"置信度 0.95"就以为"这个点 95% 是对的"。**不是。**它表示的是模型的内部把握程度，与真实正确率没有换算关系。一只被遮挡的手，模型可能凭"人一般手在身侧"的先验给出 0.9 的置信度，位置却完全错误。判断数据是否可靠，最终仍要**回看原视频**——这也是本章把"对照检查"写进实践步骤的原因。

◆ 骨架图：把散落的点连成结构

有了点，还需要把它们连起来，才能表达"这是一条手臂"。连接关系不是模型猜的，而是**写死在标准里的固定表**：左肩连左肘，左肘连左腕，左肩连左髋，左髋连左膝，左膝连左踝，鼻连双眼、双眼连双耳……把点当作**节点**（node）、把这些固定的连接当作**边**（edge），一帧姿态就变成了一张标准的图（graph）结构——所以整套数据在论文里常被叫作**骨架图**（skeleton graph）。

理解"图结构"为什么重要？它解释了两件事：**第一，为什么姿态数据可以被不同工具互相搬运**——只要两边的连接表一致（或者能通过一张转换表对应起来），数据就能互换，这正是 DWPose 的输出能直接替换 OpenPose 输出的原因；**第二，为什么时序建模可以做得比"逐个点滤波"更聪明**——既然数据天然是图，就可以用在图上做卷积的模型（图卷积网络，GCN）一次处理整段动作，让"手腕的抖动"参考"手肘的变化"来判断它到底是不是真的在动。

还有一层结构值得多看一眼：**连接是有方向与层级的。**在骨架图里，"肩"是"肘"的父节点，"肘"是"腕"的父节点；沿着这条线一路向上，所有点最终都汇到骨盆（髋部中心）。这种层级结构在动画学里叫**运动链**（kinematic chain），它带来两条贯穿全章的推论。**推论一：位置的误差会顺着链条传递与放大。**肩点若抖动 2 个像素，肘与腕就会跟着画出一条明显放大的弧线——这解释了为什么手腕与脚踝总是全身最"不安分"的点，也解释了为什么质检时要对末梢关节单独设规则。**推论二：骨长在模型里被假设不变。**一个人张开手臂时，肘到腕那段"骨头"不会真的变长，但算法输出的坐标会有轻微伸缩；迁移到角色身上时，这点伸缩会被放大成可见的变形——6.5 节的比例适配要处理的，正是这件事。

最后记住一个点：**髋部中心，常被当作"根节点"。**它是整张骨架的立足点，动作中的移动、转身、跳跃最终都会反映为根节点的轨迹；后文的"根节点漂移"质检项与重定向的足部锁定，都要以它为参照。

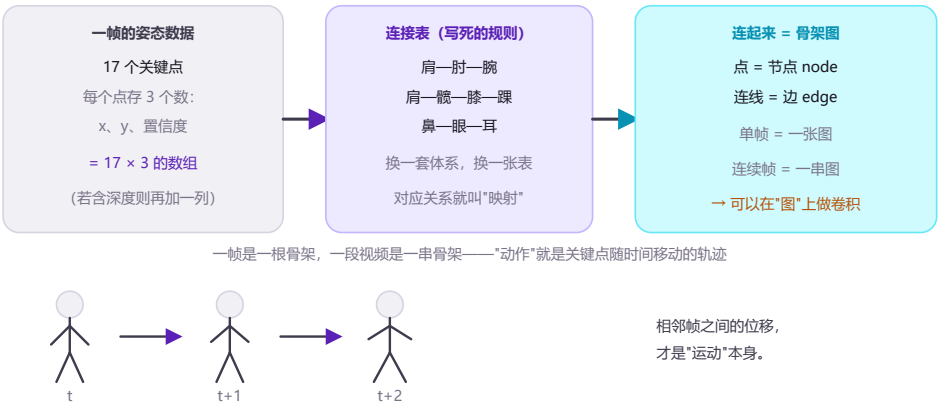


图 6-4　从一帧到一段。上半部分是数据的结构（数组 → 连接表 → 图），下半部分是同一结构在时间轴上的重复。图卷积、时序滤波、动作匹配都建立在这个结构之上。

◆ 工具定位：MediaPipe • DWPose • OpenPose

知道原理之后，选工具就变得简单了——因为每一步该用哪种工具，取决于你手上有什么、要什么。给一个极其粗略的定位口诀：要快、要能跑在手机上，用 MediaPipe；要准、尤其是手部要准，用 DWPose；要"老流程兼容"或者数据集里全是这个格式，用 OpenPose。下面这张表把三者的差别讲清楚。

表 6-3　三套主流姿态工具的能力对比（据 2026 年公开资料整理）

维度	MediaPipe Pose	DWPose	OpenPose
出品方	Google	IDEA Research	CMU
点数体系	33 点（BlazePose）	133 点（COCO-WholeBody）	135 点（身体 / 脚 / 脸 / 手）
架构思路	面向移动端的轻量模型，自带内部平滑	基于 RTMPose（CNN + Transformer），**两阶段**：先检测人，再对每个人裁剪出的区域单独跑姿态模型	**单阶段自底向上**：整图一次处理，用部位亲和场（PAF）连接关节点，无需单独的人体检测步骤
手部表现	有手部相关点，精度一般	**优势显著**，两只手都能捕捉	经常漏检手部，手指细节粗糙
精度口径	面向实时与低算力，不追求最高 AP	全身 AP 约 66.5%（COCO-WholeBody）	开创性方案，技术较老，全身 AP 明显更低
许可	开源生态，可商用	**Apache-2.0，可自由商用**	**非商业 / 学术许可，商用受限**
本课程用法	课后自测、手机端实时试验	课堂主线流程的首选提取器	作为格式与历史脉络的参照

有一个非常实用的细节：**DWPose 的输出兼容 OpenPose 格式的骨架**，两者可以在既有工作流中直接互换节点做 A/B 对比。在 ControlNet 里，DWPose 的预处理器名为 `dw_openpose_full`，社区沿用了 OpenPose 的叫法，管它叫 "DW Openpose"。这意味着你不必为了换一个更准的提取器而重建整条流水线——**换一个预处理器，其余照旧**。这背后是一个更普遍的工程经验：选择"事实标准"的格式，等于给自己留了一条退路。

> 🧪 **原理注 • 两阶段为什么对手部更友好**
>
> 单阶段自底向上的方案（OpenPose）要在整幅图里一次性找出所有人体的所有关节，人一多、一遮挡，就难免"张冠李戴"——把 A 的手腕连到 B 的身上。两阶段方案（DWPose）先框出每个人，再在这个小框里单独跑一次姿态模型——**框内的手臂相对更大、干扰更少**，手部这种"小而密集"的结构自然更准。代价是速度略慢，并且在拥挤场景中依赖检测框的准确。这个"先定位再细化"的思路，与第 2 章讲的"局部重绘"、第 4 章讲的"部件级细化"是同一种工程智慧：**把难题拆小，比硬解难题容易。**

> **🎯 要点 • 6.1 节三句话**
>
> ① 姿态估计的产物是**关键点序列**，不是图片；不同工具用不同的点数体系，转换才是常态。
> ② 模型先输出**热图**（位置的可能性地图），再从峰值取坐标；**置信度**回答"模型有多确定"，不回答"对不对"。
> ③ 关键点靠**固定的连接表**连成骨架图——这是数据能跨工具搬运、也能被图卷积统一处理的原因。

6.2 三维姿态与人体模型：从"投影"回到"身体"

二维关键点已经能做很多事了：控制生成图里人物的姿势、做动作比对、给别人演示"这个动作怎么做才对"。但它有一个硬伤：**它只是一张纸上的剪影。**当你要让一个三维角色完成同一个动作时，问题立刻来了——画面里那个人的手，是在身体前面，还是在身体后面？他的身体是朝左 30 度，还是朝右 30 度？二维投影有无数种三维解释，而三维角色只能接受其中一种。这一节讲的就是"从投影回到身体"的三件事：深度从哪里来、人体模型是什么、以及骨架之间怎么对应。

◆ 深度：二维投影丢掉的第三个维度

三维世界里的一个点 (X, Y, Z)，投影到照片上只剩下 (x, y)。Z 去哪了？**它没有被丢掉，而是被"混进了比例里"。**同一个人，站得离相机近，在照片里就大；站得远，在照片里就小。也就是说，照片里那个人的大小同时受"真实身高"和"距离"两个因素影响，而照片本身无法告诉你哪个因素在起作用——这就叫**尺度歧义**（scale ambiguity）。

更麻烦的是关节的深度：手从身前移到身后，在正面视角的照片里，手腕的位置可能几乎没有变化（因为前后的移动被投影压扁了），但三维里这只手已经划过了半米的弧线。模型必须从"手臂看起来变短了"这种蛛丝马迹里反推出"它往画面深处伸过去了"——这是一件本质上不够确定的事。

图 6-5　尺度歧义与深度歧义。左图是相机得到的证据，右图是两份都成立的解释。单目三维姿态之所以"看起来能跑但没有那么准"，根子就在这张图上。

◆ 三角化：多一双眼睛，就多一个约束

摆脱歧义最直接的办法是**再加一台相机。**如果两个机位同时拍到同一个人，那么"这个人手腕的三维位置"必须同时满足两个相机的投影关系——两条从相机出发的视线在空间中相交，交点就是答案。这个过程叫**三角化**（triangulation）。两台相机不够就三台，约束越多，解越稳。

三角化的质量取决于两件事：**视角差异**（两台相机离得太近，"两条视线"几乎是同一条，交不出来）与**关键点质量**（某个机位没看到手，那个机位就提供不了约束）。这就引出了上一节表 6-2 里"加权三角化"的用途：每个二维点带着自己的置信度参与求解，高置信度的点说话更有分量。工程上更常见的简化做法是"单目 + 先验"：只用一台相机，但假设"人是一个符合人体比例与关节活动范围的物体"，用这个假设去补足缺失的约束——这正是下一小节要讲的人体模型。

◆ SMPL：一个能被"拧"成任何姿势的人

SMPL（Skinned Multi-Person Linear model，多人线性蒙皮模型）是三维人体领域最著名的一个"参数化人体模型"。说人话：**它是一具"可以被连续调节的人"**——你给它两组数字，它就吐出一个完整的三维人体网格（带姿态、带体型），不需要任何人手工建模。

这两组数字分别控制两件事：**姿态**与**体型**。姿态用 72 个参数表示，对应 24 个关节、每个关节 3 个旋转自由度（可以想象成"每根骨头能往三个方向拧"）；体型用 10 个形状系数（shape coefficients）表示，控制高矮、胖瘦、肩宽这类"骨架之外的肉与骨"。除此之外，模型内部还包含一个学习得到的**关节回归矩阵**（把姿态和形状映射到三维关节位置）、一套**混合权重**（每个网格顶点受哪些关节影响）以及一组**姿态相关的混合形状**（blend shapes，用来模拟抬手时肩膀肌肉的鼓起、弯腰时腹部的折叠这类"随姿势变化的形变"）。

表 6-4　SMPL 与 SMPL-X 的参数构成

项目	SMPL	SMPL-X	对应到创作里的说法
关节数	24	含面部与手部，关节更多	"身体能被拆成几节来拧"
姿态参数	72	165	24 关节 × 3 个旋转；SMPL-X 额外加了脸与手
形状系数	10	10	控制高矮胖瘦，与姿态无关
内部机制	关节回归矩阵 + 混合权重 + 姿态相关混合形状	同左，并为手 / 脸增加专门的参数与先验	"它会替你补出合理的肌肉与关节形状"
典型用法	把三角化后的三维关键点**拟合**到一张人体网格上	需要手指与表情的表演级动作	关键点 → 一个能进三维软件继续加工的"人"

SMPL 在姿态估计流水线里的角色，是**那个"兜底的人"**。三角化出来的三维关键点只是一堆散点，没有体积、没有关节连接、更没法直接驱动角色；把它们拟合到 SMPL 上，就得到了一张有拓扑、有关节、有比例的完整人体——可以直接进入第 5 章讲的那套骨骼与蒙皮体系。拟合时通常只用关键点中的一部分：比如把某个标准下三角化输出的几百个点，只取其中 24 个与 SMPL 关节对应的点去做拟合。

而"用 SMPL 把二维关键点变成三维"这条路上，有一个已经相当成熟的工程配方：**先把二维关键点经 SMPL 的关节回归转成三维，再对关节角做噪声滤波，最后用贝塞尔曲线拟合来强制运动链的平滑性。**也就是说，平滑这一步不是"事后美化"，而是三维姿态估计流水线里的标准组成部分——这一点在 6.3 节还会展开。

◆ 骨架拓扑不匹配：一个必须提前知道的坑

看到这里你可能会想：既然 SMPL 这么好，那直接把 MediaPipe 的 33 点灌进去不就行了？还真不行。问题出在一个非常具体的差异上：**BlazePose 的骨架"非常像标准的 17 点 COCO 关键点，但它缺少一根脊柱骨"**。而绝大多数虚拟角色骨架（包括 SMPL、包括第 5 章自动绑定工具生成的骨架）都含有脊柱链条。

MediaPipe 的 33 点确实可以按解剖对应关系映射到 COCO 的 17 点，工程上常见做法是把输出存成 (N, 17, 3) 的数组（N 为帧数）。但"能映射"不等于"能直接驱动"——映射之后你仍然缺脊柱、缺手指、缺脚掌细节。架在这两套体系之间的，不是一根数据线，而是一张需要你自己核对与补齐的对应表。已有工具（如 Kalidokit 一类）专门做这件事：把 MediaPipe 的面部、手部、身体关键点转换成能直接驱动三维模型的数据。它们的价值，正是把这张对应表替你写好了。

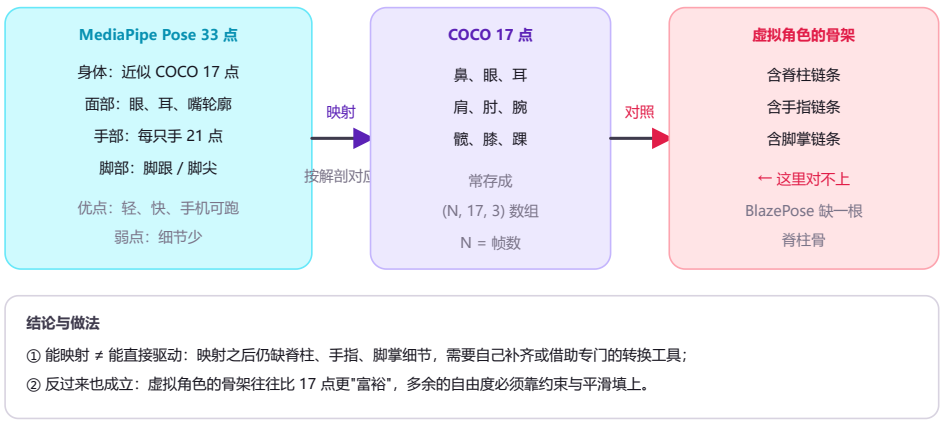


图 6-6　映射的三个环节与一处缺口。数据从"33 点"走到"虚拟角色"要经过两次换算；右边那一列不是你的敌人，而是提醒你：**任何"一键驱动"的工具，都在替你猜那些没被拍到的自由度。**

∫ 深入一格 · 为什么"深度"不能从单张图唯一确定 选学 · 数学直觉

相机投影可以写成一个简单的比例关系：设相机在原点，三维点 (X, Y, Z) 投到图像上的位置正比于 X/Z 与 Y/Z（Z 为深度）。现在反过来问：已知图像坐标 (x, y)，求 (X, Y, Z)。这个方程有**无穷多组解**——只要把 X、Y 同时乘以任意倍数 k，再把 Z 也乘以 k，投影结果一模一样。用一句话说：

$$(X, Y, Z) \text{ 与 } (kX, kY, kZ) \text{ 投影到同一个像素}$$

这就是"尺度歧义"：所有的三维解构成一条从相机出发的射线，"深度"只是这条射线上的一个自由参数。要确定它，必须引入额外的信息——

额外的信息从哪里来？三条路：**①多一个视角**（另一台相机给出独立的另一条射线，两条射线相交即解，这就是三角化）；**②物体已知尺寸**（比如"这是一颗直径 40 厘米的球"或"人的臂长在合理范围内"，即人体模型先验）；**③运动**（同一个人走两步，透视变化会暴露深度，这也是为什么"原地不动的单帧"最难恢复三维）。这三条路也解释了本章的一个实用结论：想让三维姿态稳，最便宜的办法是"多拍一点、多走两步、别在同一个机位一动不动"。

6.3 抖动、遮挡与平滑：把毛刺磨平

现在你手上有一份关键点序列了。把它画成曲线，你会看到一件让人绝望的事：**明明是一个平稳的动作，曲线却在疯狂锯齿**。手腕的位置帧帧都在上下跳动十几像素，某些帧的手直接"消失"了，还有几帧里左右手突然互换了。这不是你的素材拍得差——这是所有姿态估计的常态。这一节讲三件事：抖动从哪来、怎么把它量化、以及怎么在不"糊掉"动作的前提下把它磨平。

◆ 抖动从哪来：六个来源

原始关键点序列在视频中呈现**高频抖动与跳变**，这不是单一原因造成的。搞清来源很重要，因为不同来源要用不同办法对付，用错药反而更糟——比如把"真实的快速动作"当成抖动滤掉，人物就会变成慢动作。

表 6-5　抖动的六个来源与对应的处理思路

来源	机制	处理思路
图像噪声	传感器噪声与压缩伪影让每帧的像素略有不同，模型据此给出略有不同的位置	时序滤波（首选）；提高拍摄码率与光照可减轻
遮挡	关节被身体或道具挡住，模型只能猜；遮挡解除的瞬间位置会"跳"回来	减少遮挡输入；对遮挡区间做置信度标记，单独处理
模型敏感性	模型对微小的姿态变化过度敏感，相邻两帧的输入几乎相同却给出不同的解	时序滤波；用两端置信度都高的帧对来评估抖动
快速运动	帧间位移大，出现运动模糊、肢体拖影，模型找不到清晰边界	提高帧率拍摄；**不要把这类"真运动"连同抖动一起滤掉**
光照变化	光照影响特征的一致性，背景光斑移动会让模型误判	固定光源、避免逆光与闪烁的屏幕光
多人场景的 ID 切换	多个人物交叉时，跟踪编号会互换，导致同一编号的"人"瞬间出现在别处	只拍单人；若必须多人，事后按轨迹连续性人工分割

◆ 量化抖动：一个可以写进实验报告的公式

"抖得厉害"是主观判断，写实验报告需要数字。一个能直接引用的工程指标是**归一化抖动**：取每个关键点的相邻帧欧氏位移，算这些位移的**中位数**，再除以**躯干长度**归一化。两个细节让它更可靠：

- **用中位数而不是平均值**——平均值会被偶尔的一次大幅度跳变（比如遮挡解除的跳回）拉高，中位数对这类离群值不敏感，更能反映"常态的抖"。
- **只用两帧关键点都超过置信度阈值的连续帧对**——否则你量到的是"遮挡导致的消失与重现"，而不是抖动。

除以躯干长度是为了让数字**与拍摄景别无关**：远景里的人整体小，位移天然小；近景里人整体大，位移天然大。归一化之后，"一个人走路的镜头"和"一个人挥袖的特写"才能放在一起比较。这个指标的最大价值不是"给抖动打分"，而是给你一个对照实验的因变量——6.10 节会让你做"原始关键点 vs 平滑后"的对照，测的就是它。

◆ 四种时序滤波：延迟与平滑的取舍

滤波这件事，本质上是在做一笔交易：**用"反应的及时性"换"曲线的平滑"。**任何滤波都会引入某种形式的延迟，因为要判断"这一帧是不是噪声"，就必须参考前后若干帧——而"参考未来"就意味着"晚一点才能给出结果"。表 6-6 是四种常见方案，图 6-7 是它们的效果示意。

表 6-6　四种时序滤波方案的取舍

方案	做法直觉	优点	代价
简单移动平均	取最近若干帧的平均	实现最简单，能抑制白噪声	**延迟约等于窗口长度的一半**；有边缘效应；对快速运动表现差
指数加权平均 EMA	越近的帧权重越大，权重按指数衰减	实现简单、**延迟可控**（调一个系数即可）	对突然的真实运动仍会"拖"一下
Butterworth 双二阶级联	从频域设计，把"快抖"这类高频成分整体压低	通带平坦，专业滤波器设计工具支持	参数不直观；阶数高时延迟明显
混合 EMA + 卡尔曼	结合"惯性预测"与"测量修正"，按不确定性自适应	面向**低延迟、自适应、边缘保持**的复合策略，能在"该跟的时候跟上、该稳的时候稳住"	实现与调参最复杂；过度信任预测会"过冲"

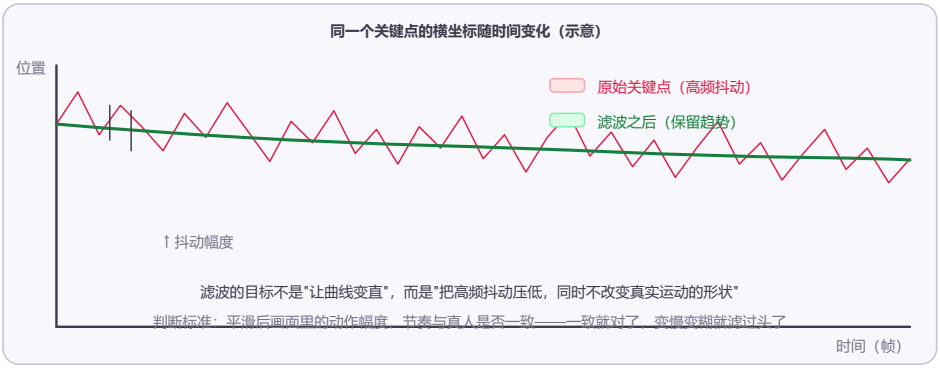


图 6-7　原始抖动与滤波后曲线。绿色曲线保留了总体趋势（位置缓慢下降），红色曲线在趋势上叠加了高频噪声。**滤波的正确目标是把红色压成绿色，而不是把绿色也压成直线。**

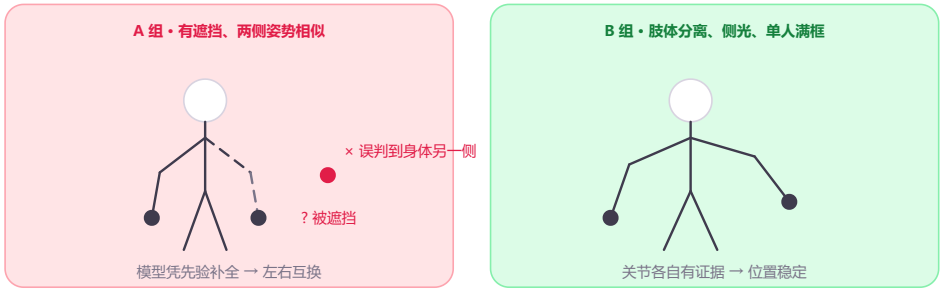
还有一个容易忽略的事实：**某些工具"已经替你平滑过一遍"了。**MediaPipe 内部带有一个轻量的卡尔曼滤波（对应参数 `smooth_landmarks=True`），但明显的抖动仍会残留，通常需要外部的时序滤波再处理一轮。这也解释了一个常见困惑——"我明明用的是同一个模型，为什么两个工具的结果一个抖一个不抖？"：它们的内部平滑设置不同。排查这类问题最快的方式，是找一个静止不动的镜头：如果连"人站着不动"都会有明显抖动，那一定是提取或平滑环节的问题。

◆ 遮挡：最硬的一块骨头

在姿态估计的所有难题里，**遮挡**（occlusion）是公认的核心挑战——甚至有专门为"预测被遮挡关键点"设计的模型架构。原因不难理解：**被遮挡的关节在画面里根本没有证据，模型只能依靠"人体结构常识"去猜。**它的典型表现有三种：

- **位置跳变**：短暂被遮挡的肢体重新出现时，关键点可能瞬间跳到另一个位置，形成肉眼可见的"抽搐"。
- **左右互换**：当右手部分被遮挡、或两侧姿势相似时，模型可能把左右镜像或交换——遮挡 + 对称，是"镜面翻转"的两个主要触发条件。
- **无中生有**：把两个相邻关节"粘连"成一个，或捏造出并不存在的关节位置，输出一个解剖上不可能的姿势。

光照变化也会影响特征的一致性，让"没被遮挡但处于阴影里"的关节同样变得不可靠。对策的优先级非常明确：**能避免遮挡就避免**（调整机位、让肢体在画面里分离、用侧光而不是大平光），**避免不了就标记**（把低置信度区间标出来，人工确认或补拍），**最后才考虑算法修复。**



处方：先改输入，再改算法。调整机位让肢体在画面里分开、用侧光勾勒轮廓、保证全身入框；
确实做不到的镜头，就在拍摄清单里标出来，单独重拍——这比事后修补省时间得多。

图 6-8　遮挡的 A / B 对照。两组的差别不在模型，而在**输入条件**。这本书反复出现同一个结论：**在生成类任务里，改输入往往比调参数的收益大得多**（第 2、3 章的"提示词改写"是同一个道理）。

◆ 丢帧之后：修复还是重拍

当某几帧或某个关节整段缺失时，你有两条路：**补**或**重来**。判断标准不是"能不能补"，而是"补出来的东西是否影响了观众会注意到的部分"。下面这张决策表可以直接当工作清单用。

表 6-7　缺失数据的处理决策表

情况	建议做法	理由
缺失 1–3 帧，两端置信度都高	**线性或样条插值补齐**	时间极短，插值误差不可见；后接滤波一起做
缺失 4–15 帧，位于动作缓慢段	插值 + 平滑；在记录里标注	慢速段的插值仍然可信，但要在技术说明里说明"该段为补全"
缺失出现在快速动作的关键瞬间	**重拍该镜头**	快速段的插值会"编"出没有的运动，观众一眼能看出节奏不对
整段动作中某个关节持续低置信度	先查原因（遮挡 / 服装 / 光线），再决定重拍或换机位	持续低置信度通常意味着采集条件系统性问题，插值解决不了
左右出现互换（镜面翻转）	**不能靠平滑修复**，必须人工标注或重拍	互换是"离散错误"，滤波只会把错误平滑得更自然，反而更难发现

> ⚠️ **一条铁律**
>
> **平滑不能修复错误，只能掩盖错误。**滤波的前提是"信号是真值 + 噪声"，而镜面翻转、穿模、关节粘连属于**系统性错误**——它们不是噪声，是错误答案。把错误答案滤得更顺滑，只会让评审和观众更难发现，等到合成阶段暴露时，返工成本已经翻了几倍。**先保证数据正确，再谈平滑。**

◆ 时序一致性：单帧都对，连起来不一定对

到这里，我们已经会检查单帧的质量了。但对动画来说，观众看到的从来不是一帧帧静像，而是一条连续的运动——**时序一致性**（temporal consistency）衡量的正是"这套动作在时间上立不立得住"：位置不跳（不闪烁、不抖动）、速度连续（没有凭空的急加速与急停）、关节关系稳定（骨长不忽长忽短）。一条数据完全可能每一帧单独看都没问题，连起来却像被人反复按了快进——这种失败在单帧检查里根本看不出来。

一个非常实用的技巧是**把数据画出来看。**任选一个关节（比如右手腕），把它的 x 坐标随帧号画成一条曲线：健康的数据是一条平滑的波浪；出现密集毛刺，那是抖动；出现阶梯状的"跳台"，往往是丢帧后被插值补出来的可疑段；出现瞬时尖峰后立刻回位，常见于遮挡解除的一瞬——关键点"啪"地跳回正确位置。很多同学在电脑前把视频反复看十遍也找不出问题，把曲线一画，第一眼就暴露了。数字不撒谎，但要画出来才看得懂。这也是 6.10 节对照实验里"画一条对比曲线"这个要求的由来。

◆ 一份姿态数据的质检单

拍摄与提取完成后，先别急着进下一步。用下面这张清单过一遍，可以省掉后面大量返工。这些指标也是 6.10 节对照实验的记录项。

表 6-8　姿态数据质检单（可打印张贴）

检查项	怎么看	合格线（经验值）
检出率	置信度高于约 0.3 的关键点所占比例	核心关节（肩、肘、髋、膝）应接近全程可检出
平均置信度	全部关键点置信度的平均值	整体偏低说明拍摄条件有问题，优先补拍
缺失帧率	关键点数量不足的帧所占比例	连续缺失超过十余帧就要评估是否重拍
归一化抖动	相邻帧位移中位数 ÷ 躯干长度（只用双端高置信度的帧对）	平滑前后各记一次，用于对照实验报告
左右一致性	回看原视频，抽查"抬手 / 转身"等易混淆帧	不允许出现左右互换；出现即修复或重拍
根节点漂移	观察髋部中心是否在"人没动"时缓慢平移	静止段不应有明显漂移，否则重定向后会"自己走"
与画面的一致性	把骨架叠加回原视频看一遍	**必做**：这是唯一能发现"高置信度错误"的办法

∫ 深入一格 · 指数加权平均只有一行公式 选学 · 数学直觉

EMA 是本章最值得动手实现一遍的算法，因为它的全部内容就是一行：

$$s(t) = \alpha \cdot x(t) + (1 - \alpha) \cdot s(t-1)$$

x(t) 是这一帧的测量值，s(t) 是平滑后的输出，α 是 0 到 1 之间的系数。

读法：**新的平滑值 = 一部分新测量 + 剩下的部分旧状态。**α 越大，越相信当前这一帧（跟随快、平滑弱）；α 越小，越相信历史（平滑强、延迟大）。取 α = 1 就退化成"不平滑"，取 α → 0 就变成"几乎不动"。你可以把 α 理解为"新证据的权重"，这与第 1 章讲的"注意力是加权平均"是同一类思想：几乎所有的"平滑"与"融合"，骨子里都是在做加权平均，区别只在权重怎么定。卡尔曼滤波比它多做了两件事：让权重随"我对预测有多确定"自动变化；并显式地把"运动的惯性"（速度、加速度）写进模型。所以它在"该跟的时候跟得上、该稳的时候稳得住"这一点上更强，代价是调参更复杂。

🎯 要点 · 6.2–6.3 节三句话

① 二维投影丢掉的是深度与尺度：**单目三维姿态的"猜"，是本质性的**，多视角或人体模型先验是把它变成"算"的办法。
② SMPL 是"可被拧成任意姿势的人"，把散点关键点变成有拓扑、可驱动的人体网格；**但骨架拓扑不匹配（缺脊柱骨）必须先解决。**
③ 抖动是常态，不是失败；**先量化（归一化抖动）、再滤波（EMA/卡尔曼）、最后回看原视频确认**——顺序反了就会白忙。

6.4 姿态条件控制：让 AI 按你指定的姿势作画

前面三节，我们把真人变成了数据。现在要让数据反向生长出一个"人"。这条路的入口在第 2 章就埋好了——**ControlNet**，一种给扩散模型加"空间约束"的技术；而在姿态这件事上，最常用的约束形态就是**骨架图**。这一节讲它怎么工作、骨架图怎么做、参数怎么调，以及国内网页工具里那些被叫作"动作参考""动作模仿"的功能，背后原理究竟是什么。

◆ 为什么文字描述姿势不够用

先看一个具体的问题。你想让画面里的人"坐在江边石头上，左手扶着膝盖，右手抬起袖子挡在脸侧"——这句话对人是清楚的，对生成模型却远远不够。提示词里的"坐着""抬着手"只能粗控人物姿势：**手抬多高、肘弯多少度、身体朝哪一侧，全是随机的。**跑十次得到十个不同的姿势，其中也许只有一两次接近你的想象。更糟的是，"手臂挡住了脸"这种要求，模型可能理解成"手臂在脸旁边"，也可能理解成"手臂不在画面里"。

要精确锁定关节位置与肢体朝向，就得给模型一个"空间约束"——不是用文字说"手抬起来"，而是直接告诉它"手在这个位置"。ControlNet 的 OpenPose 模型就是干这个的：输入一张人体骨架图（关键点连线），生成结果会与参考姿势高度接近。骨架图用的是最有名的那种视觉形态——**黑色背景上的彩色火柴人**，它不是给人看的漂亮图，而是一份"关节位置清单"的可视化。

一张流程图看清姿态条件生成

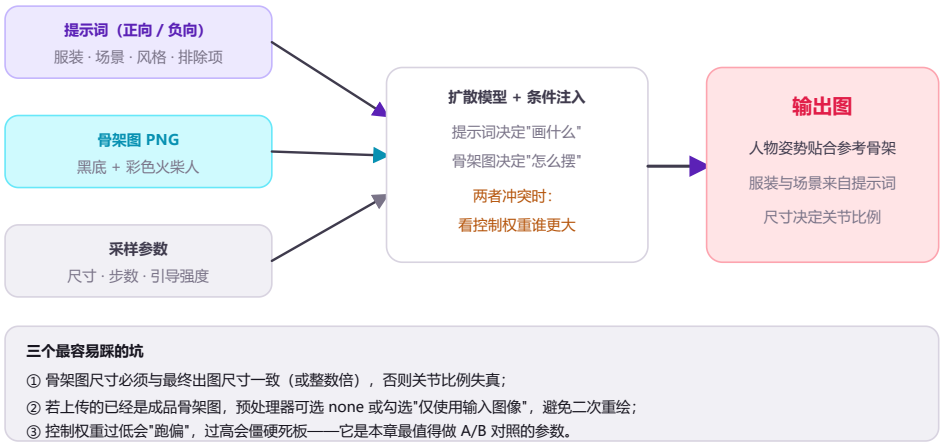


图 6-9　姿态条件生成的流程。骨架图不是"参考图"，而是**约束**：它不参与决定画的是谁、穿的什么，只决定关节在哪、肢体朝哪。**理解这一点，你就能理解为什么"换了骨架图还是同一个人"和"换了提示词还是同一个姿势"这两件事并不矛盾。**

骨架图怎么做：四步标准流程

骨架图可以"从零画"、"从照片提取"、"用软件摆"。三者的成品形态一致，难度不同。下面是被反复使用、可以直接照做的四步流程：

1. **制作骨架图。**设置画布尺寸（常见做法是 512×512，**且必须与最终出图尺寸一致**），加载内置预设、或拖拽关节点编辑姿势。注意手部若需要精细控制，要考虑到手部的 21 个子关键点。也可以直接用姿态估计工具从参考照片自动提取骨架，或者干脆手绘一个简易火柴人。

2. **导出。**保存为 PNG，要求**单通道灰度、纯黑背景、骨骼线条清晰连续、无装饰元素**，分辨率为 512×512 或其整数倍。软件通常还能同时导出 JSON 姿势数据，方便再次导入修改。

3. **在 ControlNet 中加载。**启用 ControlNet，上传骨骼图；预处理器选择 `openpose`（`openpose_full` 可同时识别手部与面部，DWPose 版为 `dw_openpose_full`）；模型选择对应的 OpenPose 控制模型；**若上传的已是骨骼图，预处理器可选 none，或勾选"仅使用输入图像"，避免自动重绘干扰。**

4. **写提示词并生成。**输入明确的主体描述与动作短语（例如 "kneeling on right knee, arms crossed over chest"），并设置采样参数。

表 6-9　骨架图的三条制作规范（决定成败的细节）

规范	要求	违反后的表现
尺寸一致	骨架图画布与出图尺寸相同，或成整数倍关系	关节比例失真——人变"长腿"或"短手"，且往往整批图都错
画面干净	纯黑背景、只有骨架线、无文字无装饰	背景被当成"内容"画进结果，出现莫名的色块与纹理
线条连续	骨骼连线清晰、粗细均匀、关节处不被截断	肢体在关节处断开，出现"两条腿不连着"或"多出一段手臂"

你可能已经发现：这三条规范里没有一条是关于"画得好不好看"的。没错——骨架图是给模型看的，不是给人看的。它越像一个标准的"技术图纸"，效果越稳定。这也是为什么很多人在这一步栽跟头：他们精心画了一张漂亮的姿势小图，结果模型理解得一塌糊涂。

参数：三个旋钮，一组参考值

姿态条件生成里真正需要调的参数不多，但每个都有明确的"过头症状"。下表是社区与教程里流传较广的参考值——**请把它们当作起点而不是标准答案**：不同模型、不同版本的最优值会漂移，本书一贯的态度是"用对照实验在你自己的工具上确认一遍"。

表 6-10　姿态条件生成的参数参考值（以你的工具实测为准）

参数	参考范围	调过头会怎样	调不足会怎样
控制权重	约 0.8–1.5；为了确保动作不变形，常用 1.2–1.5。含 DWPose 的社区工作流推荐 **0.7–0.9**	姿势僵硬、动作像被"钉住"，服装与风格被骨架线干扰	姿势跑偏，生成结果与骨架图对不上
起止步数（介入区间）	0.0–1.0 或 0.1–0.9；社区工作流常见结束步数 **0.7–0.8**	全程强约束，画面呆板、细节损失	后半程"放飞"，姿势在生成后期被改写
采样方法 / 步数 / 引导强度	如 DPM++ SDE Karras、**30–35 步**、引导强度 **6–10**	步数过高收益递减、耗时增加；引导强度过高色彩僵硬	步数过低细节不收敛；引导强度过低不听提示词
图像尺寸	**必须与输出一致**	—	不一致即关节比例失真，这是最常见的"整体跑偏"原因
多 ControlNet 组合	第一单元 openpose 控姿态（权重 1.2–1.5）；第二单元 canny 控轮廓（0.8–1.0）；手部细节可用深度图单元强化	多个高权重单元"打架"，画面撕裂	只靠骨架，细节（服装轮廓、手）不可控
负向提示词	抑制 `deformed hands`、`extra fingers`、`bad anatomy`	过度堆砌负面词会压制正常特征	手指与解剖错误率上升

◆ 网页工具里的"动作参考"：同一件事的三种皮

你大概不会去搭一套 Stable Diffusion + ControlNet。没关系——同样的能力已经被包装成了网页上的若干形态，它们在原理上都属于"用姿态作为生成条件"：

- ▣ **浏览器端骨架编辑器** · 可离线 · 免费　推荐入门
- ▢ **绘图工具的姿态插件** · 在工具内编辑姿势后直接送入生成　进阶
- ▢ **在线人偶摆姿网站** · 拖拽人偶摆姿势，导出图　轻量
- ▢ **平台集成的姿态控制** · 上传标准骨架图、在线编辑关节点　最省事

其中一类特别值得知道：**轻量、浏览器端、支持离线、免费、无需登录也无需编程的骨架编辑器**——你可以拖拽关节点、加载内置或自定义预设、整体平移姿势，再导出 PNG（黑底彩色火柴人）和 JSON 姿势数据，JSON 还能再次导入修改。它导出的骨架图与主流 ControlNet 的 OpenPose 兼容，输出可用于个人或商业作品。这类工具的价值在于"闭环"：导出 → 不满意 → 导回 → 改 → 再导出，姿势可以反复微调到满意为止。

另一类是国内平台把 ControlNet 直接集成进产品，支持上传标准骨架图、在线编辑关节点。对非技术背景的同学来说，这是"够用且省事"的路径：**能力一样，代价是可控性略低、参数被平台藏起来了**。至于平台上的"动作参考""动作模仿"等功能，它们的完整原理会在 6.6 节讲——**那已经不只是"一张图的姿势"，而是"一整段动作的时间序列"**。

最后补一句边界，它能帮你少走弯路：**姿态条件管的只是"身体怎么摆"，不管"长什么样、穿什么、在哪里、什么风格、什么情绪"**。它不控制五官长相、不控制服装材质、不控制背景与光线、不控制画面风格——这些都归提示词、参考图（第 2 章讲的角色一致性工具）与随机种子管。反过来，它也**不能替你修正解剖错误**：如果骨架图本身就是错的（比如手部关键点粘连），生成结果会忠实地把错误"画"出来。骨架图是"逐字的稿子"，模型是"照稿誊写的人"——稿子上写错了字，誊得多漂亮也还是错字。

🩺 **案例 6-A** § 捣衣仕女：从 9 秒演示到一张定妆图

背景：春江小组要把"捣衣"这个动作放进短片的一个关键镜头。小月想要一张"月色江边、仕女举杵捣衣"的画面，前后试了三轮提示词，人物的姿势每次都不同：有的像在投掷，有的像在打伞。

做法：小舟用手机拍了自己做捣衣动作的照片（单人全身、侧光、肢体分离），用姿态估计提取关键点，导出黑底彩色火柴人骨架图；把骨架图作为姿态条件送进生成，提示词只负责"汉服、月色、江边、青蓝调、水墨动画风格"这类与姿势无关的部分；控制权重从 1.0 起做对照，最后落在 1.3 附近；尺寸统一为 16:9 的输出比例，骨架图画布按同比例制作。

结果：第四轮生成时，人物的手臂角度、躯干倾斜与真人几乎一致，杵的高度与"落下"的动势也对得上。小月在这个基础上做局部重绘，把手部与袖口又修了一轮。

复盘：①把"姿势"和"内容"彻底分开是这次成功的关键——姿势交给骨架图，内容交给提示词，各司其职；②第一轮的失败并不是提示词写错了，而是要求模型用文字完成它不擅长的事；③真人演示比"摆人偶"更自然，尤其是手腕与肩部的细微角度。

6.5 重定向：把动作交给另一个身体

图像路线的终点是一张张静止的画面。而一旦你要让角色**连续地演一段戏**，就必须把动作交给三维角色——这时你面对的是本章最核心、也最容易被低估的一步：**动作重定向**（motion retargeting）。第 5 章 §5.7 已经给过定义与原理（四类前置条件、"旋转可搬、位置要换算"的量纲直觉）：把动画从一个角色/骨架迁移到另一个。**本节不重复那部分，只做一件事——把它拆成可以逐关验收的三道关。**这里我们要把它拆开，看清它到底难在哪——因为迁移失败的绝大多数原因，都能归结为下面这三道关。先从一道"看不见的关"说起：骨架之间到底是怎么对上号的？

◆ 第一关 · 骨架映射：先给关节起对名字

迁移动作的第一步不是数学，而是**命名。**源骨架（你提取出来的真人姿态）和目标骨架（第 5 章的角色）各自有一套关节名称，必须建立一张对应表，才能知道"源的左肘"该驱动"目标的哪一节骨头"。这张表如果没有，迁移就无从谈起。

这也解释了工程界一个看起来不起眼、实际价值巨大的事实：**重定向的质量取决于源骨架与目标骨架的命名规范与比例相似度。**当一个平台的骨骼采用广泛兼容的命名（例如与主流动捕库一致的命名规范），跨工具重定向就变成了一种"通用插头"——你可以把从 A 工具绑定的角色，拿到 B 工具里播 C 工具的动作库。前面第 5 章介绍过的若干自动绑定工具都特意采用了这种兼容命名，原因就在这里。

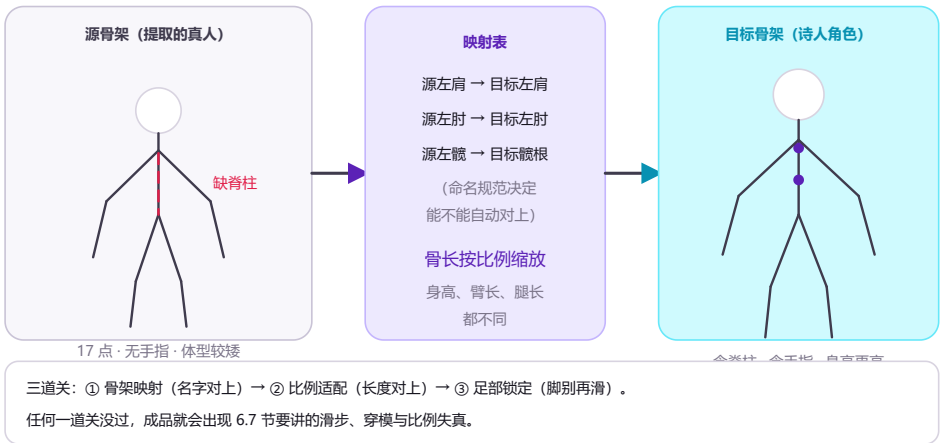


图 6-10　重定向的三道关。左边缺一根脊柱、右边多一根脊柱——多出来的自由度不会自动"知道"该怎么动，必须由算法或人工补上。**"映射表"看起来枯燥，却是整条链路上最值钱的一张纸。**

◆ 第二关 · 比例适配：一米六的动作，一米八的身体

假设映射表建好了，新的问题立刻出现：**源和目标的身材不一样。**真人 1.6 米，角色 1.8 米；真人的腿占身高比例更小，角色的手臂更长。如果把关节旋转角度原样复制过去，会得到什么？——动作的"形"对了，但"势"错了：挥袖的幅度显得过大、迈步的步幅显得夸张，看起来像是动作"被放大"了。

比例适配要做的，是把动作从"绝对位移"翻译成"相对比例"。核心思路是**归一化**：先用源骨架的尺度把动作标准化（比如把"手腕移动了 0.3 米"换算成"手腕移动了 0.2 个臂长"），再用目标骨架的尺度还原（"0.2 个臂长"× 目标自己的臂长）。这样，同一个动作在矮个子与高个子身上做出来，"看起来的幅度"是一致的。

但归一化会带来一个副作用：**脚的位置会变**。手臂短一点长一点，观众未必看得出；但如果脚的位置被改了，观众会立刻看到"这个人在冰面上滑"。于是就有了第三关。

把这件事说白了：**重定向传输的不是"关节角度"，而是"动作的意图"**——"手臂抬起到与肩同高"是意图，"肘关节旋转 47 度"只是实现方式。同一句话由不同身材的人来执行，关节角度本就应该不同。反过来，这也是判断重定向好坏的一条经验标准：看结果像不像"同一个人在用另一种身材做同一件事"，而不是看骨骼参数是否逐一对齐。工程实现上，常用做法是先做一次全身尺度的对齐（让两副骨架"身高一致"再搬运），再逐骨骼修正；对差异特别大的目标（比如人迁移到夸张的卡通体型），比例相似度的下降会让"形似"越来越难保住，此时宁可牺牲关节角度的精确，也要保住重心高度、步幅节奏这类"一看就知道像不像"的特征。

◆ 第三关 · 足部锁定：不许滑

足部锁定（foot lock / foot planting）指的是：当一只脚接触地面并且应该"踩住"时，强制它在世界坐标里保持不动，让身体从它上面"走"过去。它解决的是迁移里最刺眼的瑕疵——**滑步**（foot skating）。滑步的观感非常糟糕：角色像是在冰面上被拖着走，或者像一段"没配好音"的舞蹈。

这里有一个可以推广到视频生成的判断标准：脚该不该固定，取决于动作的语义。站立、踏步、踢腿、接触地面时，应当开启足部锁定；而跳跃、滑行、飞行这类"脚本来就该自由移动"的动作，则要关闭它。也就是说，足部锁定不是"永远更真实"的开关，而是**一个需要按动作类型判断的语义开关**——把它当成万能的修复手段，反而会把跳跃变成"粘在地上"。

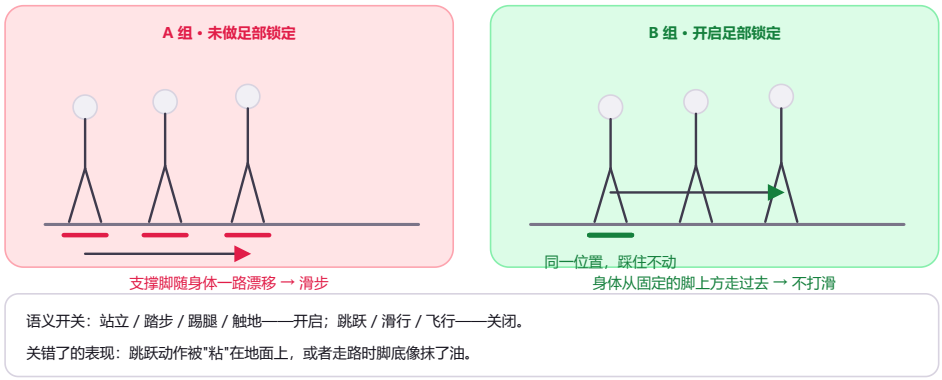


图 6-11　足部锁定的原理与边界。两组画面的身体位移完全相同，差别只在支撑脚是否被"踩住"。右下角那两行是本章最实用的一条经验——**没有万能的开关，只有按语义做出的判断。**

◆ 导出与落地：回到第 4 章的那张检查清单

重定向做完，还没结束。把带动画的三维角色送进引擎或交给下一环节之前，还有几个"跨章衔接"的问题要处理——它们全部来自第 4 章的导出检查清单，这里只做提醒，细节请回看那一章：

- **烘焙动画**：如果骨架里用了约束或反向运动学（IK），应当先"Bake Animation"把它烘焙成逐帧的正向运动学（FK）数据，否则引擎无法正确重现。
- **尺度与轴向**：不同引擎的单位与朝上轴不同（例如有的用米、有的用厘米），导入后"放大一百倍"或"横着躺下"十有八九是这里的问题，导出前把变换应用干净（缩放读作 1）。
- **Pivot（原点）**：角色的原点应放在逻辑位置（例如脚底或颈根），否则放进场景后"怎么摆都不对"。
- **面数与蒙皮质量**：形变区的边环够不够、权重有没有"渗漏"，会直接决定动作迁移后的形变是否撕裂——这是第 5 章"骨骼与蒙皮"那一节的内容。

◆ 不想自己拍？还有三条"现成动作"的路

自己拍动作是本章的主线，但工程上有更省事的替代路径，值得知道它们各自的位置：

表 6-11　三条"现成动作"的来源路径

路径	怎么用	适合什么	要注意什么
公开动捕数据集	下载公开的动作捕捉数据（含走路、体育、日常活动等大类），重定向到自己的角色	需要大量"通用日常动作"的场合	原始数据常无引擎骨骼绑定、无循环裁剪、无脚部固定清理，需要自己做转换与重定向；部分资源的手指数据实际未采集；使用须遵守其许可与致谢要求
平台动作库	从动捕动作库中挑选片段，直接用于原型制作	快速验证、占位、课堂练习	片段通常已预解算、循环就绪、引擎兼容；但仍需重定向到你的角色，且覆盖面有限、需细读商业许可条款
文生动作	用文字描述生成三维人体动作（例如"挥手、转身、缓慢起舞"）	没有合适素材、需要快速构想动作时	第 5 章介绍过的文生动作模型对非人形角色、情绪与外观属性、多人动作、原地循环等有明确不支持项；生成的骨架通常也需要重定向到自己角色

三者的共同点很值得记住：无论动作从哪来，最后都要经过"重定向到你的角色"这一步。这也就意味着，本章讲的三道关——映射、比例、足锁——是绕不过去的基本功。别人替你做好了动作，但没人能替你把它装到你的角色身上。

> **🎯 要点 • 6.4–6.5 节三句话**
>
> ① 姿态条件生成的本质是**分工**：文字管"画什么"，骨架管"怎么摆"；冲突时看控制权重。
> ② 重定向三道关：**骨架映射（名字）、比例适配（尺寸）、足部锁定（语义）**——顺序不能乱。
> ③ 动作可以从公开数据集、动作库或文生动作三种途径获得，但**"装到自己的角色上"永远是你自己的活。**

6.6 "参考动作驱动角色"的现状：一条视频，一次模仿

到目前为止，我们做的都是"一帧一帧"的事：一帧骨架 → 一张图；一串骨架 → 一段角色动画。但如果把这件事整体交给一个模型呢？**给它一段参考动作视频和一张角色图，它直接吐出一段"角色在做这个动作"的视频**——这就是"参考动作驱动角色"这条路线，也是过去两三年视频生成领域最热闹的方向之一。学这一节的目的不是让你记住哪个模型更强，而是学会判断一条技术路线"现在能不能用、能不能商用、适不适合你"。

◆ 源流：AnimateAnyone 定下的三段式

这条路线有一个被反复引用的技术源头。**AnimateAnyone**（2023 年 11 月公开论文）确立了后来被大量实现沿用的设计：**用一个"参考网络"注入外观**（保证角色还是那个角色）、**用一个"姿态引导器"把动作条件化**（保证动作还是那个动作）、**再加一层时序模块**（保证帧与帧之间不闪烁、不漂移）。这三件套解决的是同一件事的三个侧面：**外观、动作、时间一致性**。

但这里有一个非常现实、也非常值得作为"媒介素养"案例来讲的事实：**AnimateAnyone 的代码和训练权重均未公开。**它的代码仓库虽然有可观的关注度，但基本只有说明文档、许可证和演示，没有可运行的东西。后续还有 Animate Anyone 2，同样只有论文与演示。"论文里看到的能力"与"你今天能下载到的东西"之间存在一条鸿沟——这条鸿沟在本课程里必须被明确指出来。这也是本书反复强调"以官方页面为准"的原因。

◆ 开源侧：好看的路线，与难缠的许可证

开源生态里最接近"一张图 + 一段姿态序列 → 角色动画视频"的代表是 **UniAnimate**（由高校与产业团队合作提出）。它的输入很直白：**一张参考人物图像 + 一段目标姿态序列**，输出跳舞/角色动画视频。技术上有两点值得知道：它把参考图像、姿态引导与噪声视频映射到同一个特征空间，**不再依赖额外的参考网络**；并且支持首帧条件输入，可以迭代生成长度可观的视频（论文口径为最长约一分钟）。它还探索了用更省算力的时序建模方式来替代计算量大的时序 Transformer。

但把话说全：**UniAnimate 的权重虽然公开，许可证为非商用。**它的显存需求以"32 帧 768×512"为例约 12GB，整体约 12–21GB；社区对它的支持属于非官方性质。还有一个基于更强视频基座、用轻量微调方式得到的版本，分辨率更高但显存需求也水涨船高，**许可证不明确，商用需自行确认。**更值得注意的是，有对比研究指出：这个版本能很好地保留角色外观，**但动作上可能明显偏离目标动作**——**外观像、动作不像，是这条路线很典型的一种失败。**

那么开源侧有没有"真能商用"的选择？有：一篇 2026 年的比较文章给出的结论是，开源方案中**实际可商用的是 Wan 2.2 Animate（Apache-2.0 许可）**；MimicMotion 质量较好；UniAnimate 省显存但非商用。请注意这类比较文章的时效性——许可证是会变的，三个月以后你读到的结论可能就不同了。

◆ 网页侧：真正能"今天就试"的入口

对通识课的同学来说，最有价值的不是论文，而是"现在打开浏览器就能用"的路径。在检索范围内，国内网页端明确对应"上传参考动作视频驱动角色"能力的平台是**可灵 Kling 的 Motion Control**，它提供了两代主要版本：

- **较新的那一代**：网页 / App 上传**参考动作视频**和**角色图像**，也可以从动作库中选择动作；还可以把面部元素绑定起来，用以提升复杂多角度、长时间动作下的面部一致性。官方把它定位在"电影化表演、高精度动作捕捉、娱乐场景"。
- **上一代**：上传要模仿的角色动作视频，或从动作库中选择；再添加角色图像，**角色图与视频中人物的比例最好匹配。**默认"角色朝向匹配视频"，也可以选择"角色朝向匹配图像"（后者支持镜头运动）。上传的动作视频支持 **3–30 秒**，生成视频的时长通常与上传视频一致；**如果动作复杂或过快，可能只抽取有效的连续动作，生成结果会短于原视频。**

计费按秒计算（以平台当前定价为准，两代产品分专业版与标准版，价格不同），这意味着一件事：时长就是成本。这也解释了为什么"参考视频控制在 3–30 秒、动作不过快"既是技术要求，也是经济要求。

顺带建立一种习惯：读这类工具资料时，**先看日期、再看结论。**以这条产品线为例，从最早上线到本章编写时，它大致经历了"初版 → 中版本迭代并强化一致性 → 引入原生音频 → 推出旗舰版本、支持多镜头与 4K"这样一条快速演进的轨迹——平均每隔几个月就有一轮功能与定价的更新。**一个功能昨天没有、今天上线，或者昨天免费、今天收费，都是常态。**所以本书给的是"检索与判断的方法"，而不是"功能清单"：任何功能细节，请以你使用时官方页面的最新说明为准；写进报告时，注明查证日期。

另一个必须诚实交代的点：**有些平台功能在本书编写时无法核实。**例如有同学会问"某国内视频平台有没有'动作参考'功能"——在检索范围内我们没有获得该功能存在的直接证据。可以证实的是，该平台在 2026 年 8 月上线了三维软件的插件，并通过其新一代视频模型打通三维资产流程，其中包含"三维白模预览 / 白模参考"能力（先定角色位置、相机路径、空间尺度与动作节奏，再由模型补材质、灯光与环境）。**本书宁可说"截至编写时未获证实"，也不替平台编一个功能**——这也是你写技术说明时应该有的态度。

由此可以总结出一套"面对陌生平台时的试跑法"，四步即可判断一个动作迁移功能值不值得用：**第一步，看输入约束**——它接受什么（参考视频？动作库？还是只有骨架图），时长限制是多少，是否支持"角色图与视频人物比例匹配"的提示；**第二步，看输出与参数**——能否设置朝向匹配方式、是否有足部锁定等地板接触选项、生成时长是否可预期；**第三步，跑一个 5 秒的测试用例**——就用实践一拍的素材，看结果里"手指是否变形、脚下是否打滑、比例是否失真"这三项，十分钟就能得出体感结论；**第四步，查许可与计费**——生成内容能否商用、按什么计费、是否需要标注。四步走完，你对它的判断会比读十篇测评更靠得住。

图 6-12　三条路线的定位。本章的主线走第三条：**把过程拆开、每一步都留下记录**——这既是为了可控，也是为了让你的期末"技术说明"有东西可写。第一、二条路线则各自对应"快"与"深"。

◆ 整体性失败：两个输入源的冲突

这条路线最常见的"整体性失败"，不是某个关节画错了，而是**两个输入源互相冲突**：角色静态图负责外观，参考视频负责动作，当两者的差异过大时，**没有任何开关能完全修复。**一个胖角色去做芭蕾舞者的动作、一个穿长袍的角色去做需要大幅开腿的动作，都会撞上这堵墙。

实践中可以把这类失败拆成五个常见原因，每一个都有对应的排查动作：

1. **参考视频时长不当。**太短（短于约 3 秒）模型缺乏足够数据锁定动作模式；太长（超过约 30 秒）容易丢失主体一致性。**做法：把参考切成 3–30 秒的完整动作单元。**

2. **模型渲染能力超限。**一次生成十秒以上的长镜头容易丢细节、崩动作。**做法：分镜生成，后期拼接。**
3. **模式选择错误。**部分平台必须使用高品质模式，标准模式无法调用运动学约束模块；不加"固定机位"时，运镜会放大失真甚至导致肢体崩解。**做法：读清模式说明，写明机位。**
4. **中途换模型。**不同模型的潜空间偏移、时间建模逻辑与运动先验都不同，换来换去会出现动作断裂、角色形变、场景漂移。**做法：一个镜头内不换模型；跨模型沿用同一段参考视频做对照。**
5. **把一次生成当成品。**有记录的实践显示，**平均每个可用镜头需要生成约 3 次，仅约 25% 的生成片段进入最终剪辑。做法：把"生成次数"写进项目计划，而不是当作失败。**

> 📌 **路线图 · 这一节该怎么用**
>
> 你不需要记住任何版本号，需要带走的是四个问题：**①这件事我今天能跑起来吗？②许可证允许我商用/公开展示吗？③失败时我从哪里查？④它是"整体一段"还是"可拆解"的？**第 ⑤个问题——**这条路线的成果能不能进入第 8 章的智能体工作流**——会在最后一章揭晓：能被脚本调用、能被明确参数化的步骤，才容易被自动化。

6.7 失败案例诊所

这一节是本章的"急诊室"。前面所有的原理，在这里都要变成"看到症状能想到成因、想到成因能开出方子"。先讲三条贯穿性的根因，再看三个具体病例。

◆ 先记住三条通用根因

1. **模型没有持久记忆、没有三维空间、没有物理模拟。**AI 视频本质上是逐帧的统计采样，微小变化会累积成漂移、漂浮、克隆等伪影——脚脱离地面、角色瞬间位移，都源于此。
2. **你没说的部分，模型会"脑补"。**未说明的关节、未交代的空间关系，模型会用默认的骨骼模板或猜测补全——"无中生有"的关节与"多出来的一段手臂"就是这么来的。
3. **参数与模式会制造伪影。**运动幅度类参数调得过高会让关节高频抖动、或"强行插入"你没要求的动作；某些"自动增强"开关会造成关节二次拉伸。

在此基础上，再给你一条排查的顺序，能省下大量时间：**先查输入，再查参数，最后才考虑换工具。**先看参考视频与角色图是否满足规范（时长、单人、全身、尺度一致、无明显遮挡）——本章绝大多数案例的"真凶"都在这一步；其次是参数与模式（控制强度、运动幅度、是否高品质模式、是否开了不该开的自动增强）；只有在输入与参数都无懈可击时，才值得怀疑工具本身能力不足。**顺序颠倒的典型症状是：换了一圈工具，问题原样复现**——因为它从一开始就跟着你。

> 🩺 **失败诊所 F6-1** 被"想象"出来的手：遮挡误判导致的手部畸变
>
> **症状：**生成结果里，人物举手挡在脸侧，但手指粘连成一团、多出一根手指，或者手腕的位置与手臂对不上——像是手腕"悬"在半空。参考骨架图上明明标着手的位置。
>
> **成因：**三层原因叠在一起。① **提取层**：参考视频里手部被衣袖或头部遮挡，姿态估计只能凭先验猜，属于"高置信度的错误"；手部是 21 点的密集结构，一旦漏检，剩下的点会连成解剖上不可能的姿势。② **条件层**：骨架图本身缺了手指细节（17 点体系根本没有手指），模型只好自己"脑补"手——于是捏造出不存在的指节。③ **生成层**：手掌与手指在画面中占比小、结构密集，属于生成模型最不擅长的"局部解剖"区域。
>
> **处方：**① 提取时改用**手部表现更好的姿态工具**（含手部 21 点的体系），必要时让手在拍摄时不被遮挡；② 骨架图若需手部控制，采用支持手部关键点的预处理模式；③ 负向提示词抑制手部相关错误（`deformed hands`、`extra fingers`、`bad anatomy` 等）；④ 局部畸形用**只勾选手部区域的局部重绘**修复——只修手，不要选全身，否则会引起整体风格偏移；⑤ 若预算允许，把手的设计"做减法"：让角色袖子遮住手，是古装题材里最经济也最合理的规避方式。

🩺 **失败诊所 F6-2** 冰面上的舞者：迁移后的滑步

症状：角色在动作迁移后脚下"打滑"——支撑脚明明应该踩住地面，却在画面里缓缓位移；或者走路时脚掌像抹了油，一步三滑。动作本身看起来是对的，但"站不住"。

成因：① **未做足部锁定**：重定向只搬运关节旋转，脚的世界坐标随之漂移；② **参考片段本身不干净**：镜头抖动会被误读为角色运动，模型还会自行填补缺口；③ **模型没有物理模拟**：脚与地面的接触关系不是"算"出来的，而是"画"出来的，自然无法保证踩住；④ **比例适配的副作用**：腿长改变后脚的位置被顺带改了，而没有任何机制把它拉回来。

处方：① 按动作语义开启足部锁定——**站立、踏步、踢腿、接触地面时开启；跳跃、滑行、飞行时关闭**；② 参考视频必须**相机稳定、尺寸一致**，必要时用三脚架与固定机位；③ 拍摄前给支撑脚留出"踩稳"的时间（这些帧是后期锁定的锚点）；④ 拍摄时避免让脚被裙子、长袍或道具完全遮住——看不见的接触点无法锁定；⑤ 如果滑步出现在三维路线，检查重定向后是否做了接触约束与根节点修正。

🩺 **失败诊所 F6-3** 被撑破的袖子：比例失真与穿模

症状：迁移到角色身上后，动作幅度显得夸张或别扭——挥袖的弧线过大、迈步像跨栏；严重时手臂直接穿透身体的袍子、道具穿进身体。人物的"比例"看起来与角色设定不符。

成因：① **首帧角色图比例失衡**会让模型脑补出错误的解剖结构；② **参考视频与角色图的尺度不一致**，模型只能强行拉伸；③ **未做骨骼姿态约束、缺乏物理碰撞逻辑**，模型不懂人体力学；④ **大角度腾空、高速旋转、复杂肢体缠绕**这类极端动作超出渲染能力，模型会用默认骨骼模板补全；⑤ **并发动作**（比如"一边走一边大幅挥手"）会让约束互相打架。

处方：① 角色图要**全身完整、含双手双脚、服装结构清晰**；② 参考视频要**单人全身、相机稳定、尺度一致、肢体分离清晰**；③ 明确角色、道具、场景的空间边界，开启碰撞相关选项；④ **不要设计并发动作**，一个镜头只做一件事；⑤ 极端动作拆成两段，或用机位与景别规避（衣袖遮住手、侧面机位、半身景别）；⑥ 运动幅度相关参数宁小勿大——幅度参数过高会直接引发手指抽搐与布料穿模。

图 6-13　四类典型失败模式速查。把它们贴在工作台前，比记参数更管用。**"改输入优先于调参数"是这四个病例共同指向的结论。**

表 6-12　动作迁移失败模式总表（速查版）

症状	直接成因	一句话处方
滑步	未做足部锁定；片段杂乱；镜头抖动被误读为运动；模型无物理模拟	按语义开足部锁定；相机稳定、尺寸一致
穿模	无骨骼姿态约束与碰撞逻辑；极端动作超限；并发动作	开碰撞；一个镜头只做一件事；避开极端动作
手指畸变	手部遮挡与比例失衡引发模型捏造指节	换更准的提取器；只修手的局部重绘；负向提示词
镜面翻转	单侧被遮挡 + 两侧姿势相似，是主要触发条件	避免"对称姿势 + 遮挡"的输入组合；侧光与取景有助缓解
比例失真	首帧比例失衡；参考与角色尺度不一致	角色图全身完整；参考单人全身、尺度一致
抖动	逐帧采样缺时序一致性；滤波不当	时序滤波（EMA / 卡尔曼 / Butterworth）；降低运动幅度参数
整体模仿失败	外观与动作两个输入源冲突；时长不当；模式错；中途换模型	两个输入尽量同质；参考 3–30 秒；用高品质模式；一个镜头不换模型

图 6-14　要被「拆成坐标」的那个动作。本章要做的，就是把这样一个动作变成一串可以计算的关键点：先拍摄（或取用）这段动作，再用姿态估计得出骨架序列，最后迁移到第 5 章的角色身上。先看清「人做了什么」，再谈「机器读到了什么」。 AI 生成

6.8 案例教学：从舞蹈教室到屏幕

原理与失败都讲完了，现在看三个真实的"怎么把它做成"的过程。第一个是本书的主线案例，另两个来自舞蹈教学与运动技术分析这两类把姿态估计用得很实的场景。

🩺 案例 6-B § 捣衣与拂袖：春江小组的动作迁移全流程

背景：短片《春江花月夜》需要三个动作段落：开头"江畔谁人初见月"里的**拂袖回望**、中段"玉户帘中卷不去"的**捣衣**、结尾"落月摇情满江树"的**抬手远望**。小舟决定真人拍摄 + 姿态迁移，理由是"三个人自己演的比文生动作更有味道，也更容易解释技术过程"。

做法：分四步走。**第一步，定拍摄规范**——教室白墙前、单人全身入框、三脚架固定机位、侧上方打一盏灯做出轮廓、穿有清晰肩线的练功服（避免宽大袖子遮住手肘）；每次拍摄前先做一个"举手、张手、踩实"的标定动作，用于检查手部与脚部能否被稳定检出。**第二步，拍三段各 8–15 秒**（落在参考时长区间内），每段做 3 次取最好的一条；拍完立刻回放骨架叠加图，发现拂袖段右手被袖子遮住，当场重拍。**第三步，双路线并行**——图像侧用骨架图做姿态条件生成（用于短片的定格镜头），三维侧把姿态重定向到第 5 章的诗人角色（用于需要镜头运动的段落）。**第四步，质检与记录**——用表 6-8 的清单过一遍，把平滑前后的归一化抖动都记进实验记录。

结果：三段动作总共生成约 20 个候选片段，最终采用 5 个；三维侧的诗人角色在"捣衣"段出现了轻微滑步，开启足部锁定后解决；图像侧的定妆图在控制权重 1.3 时姿势最贴合，1.0 时手部开始跑偏。

复盘：①拍摄规范花的十分钟，省掉了后面两小时的返工——遮挡那一次如果没在教室发现，会在生成阶段以"手指畸变"的形式暴露，排查成本高出数倍；②"三维 + 图像"双路线共用同一份姿态数据，是有意为之：这样两条产线的一致性天然更好，也为第 8 章的合成留了接口；③把"生成次数"写进计划后，小组不再把 3 次不出片当成失败，心态与进度都稳了下来。

图 6-15　主线案例的动作迁移时间线。五个步骤里，只有第三步是"技术"，其余四步都是"规范与判断"。**这条经验对整门课都成立。**

🩺 案例 6-C § 把一支非遗舞蹈"拆"给所有人看

背景：某高校舞蹈社想把一支地方非遗舞蹈做成线上教学与传播素材。传统做法的瓶颈是：请传承人录一遍完整舞蹈容易，但"分解教学"极难——讲到"第二拍手腕要往里收"，光靠语言和示范，学员还是看不明白。

做法：社团做了三层工作。① **录制规范化**：请传承人在固定机位、均匀光照下分段录制，每个动作单元单独录一条，并请她亲口讲解动作要领（这些讲解成为教学文案的原始素材）；② **姿态提取与可视化**：用姿态估计提取关键点，把"手腕往内收"这件事翻译成"手腕与手肘夹角变小、前臂相对躯干旋转"的量化描述，做成关键帧对比图；③ **动作迁移做"风格对照"**：把提取出的动作迁移到卡通形象上，做了一版"动作示范动画"，让年轻观众先被形象吸引、再看真人讲解。

结果：分解教学的可理解度明显提升——学员反馈"看关键帧对比图，比看十遍视频有用"。迁移到卡通形象上的示范动画被用作预告片，也带来了更多报名。

复盘：①姿态数据在这里的价值不是"替换真人"，而是"解释真人"——它把难以言说的动作要领变成了可测量的量；②做迁移前，社团先取得了传承人的书面授权，并明确约定"形象仅用于本教学项目、标注 AI 生成"，这也让传承人更愿意配合（相关做法见 6.11 节）；③真人影像始终是主教材，AI 只是辅助理解——这个主次关系不能倒过来。

🩺 **案例 6-D** § 跑姿诊断：一个人也能做的"动作实验室"

背景：一位同学在校运会前想改进自己的短跑起跑姿势，但学校没有动作分析实验室，教练也只能凭经验看。

做法：他用手机在同一条跑道、同一个机位（侧向 90 度、三脚架固定）拍下自己的起跑动作，用姿态提取工具跑一遍，然后把关键点数据画成曲线：膝角随时间的变化、重心相对支撑点的高度、两手摆动的对称性。同一段动作拍三次，比较三次的关键点曲线差异有多大——这一步很关键，它把"我今天发挥得好不好"变成了"我的动作有多稳定"。

结果：他把视频与曲线一起交给教练，教练据此指出"起跑前两步躯干抬得过快"。改进后再拍一次，曲线上的躯干角度变化明显。整个过程没有用到任何专业设备。

复盘：①姿态估计最被低估的用途，是"把主观印象变成可比较的数字"；②固定机位在这里比什么都重要——机位一变，两次拍摄的数据就没法比了，这与本章反复强调的"尺度一致"是同一条原则；③这类"运动技术诊断"是姿态估计在体育与康复领域的真实应用方向，也是本课程知识的外溢价值之一。

6.9 关键概念表

这一章的名词密度是全书最高的几章之一。下表把本章出现过的核心术语集中收拢，配上英文与"一句话解释"——建议在动手实践时把它当查词表用：看不懂某个步骤在做什么时，先回来查这个词到底解决什么问题。

表 6-13　本章关键概念速查表

术语	英文	一句话解释
姿态估计	pose estimation	从图像或视频中定位人体的关键点（关节、五官），把"动作"变成一串坐标。
关键点	keypoint / landmark	人体上被算法定位的点位，如鼻、肩、肘、腕、膝、踝；常用体系有 17 点、33 点、133 点等。
骨架图	skeleton / stick figure	把关键点用线段连起来的图像，是"姿态条件"最常用的输入形式。
置信度	confidence / visibility / score	模型对"这个点在哪"的把握程度；不同架构下含义不同（热图峰值、分类概率等），用法主要是过滤与加权。
热图	heatmap	每个关键点输出的一张"概率分布图"，峰值位置即点位，峰值高度即置信度的来源之一。
自顶向下 / 自底向上	top-down / bottom-up	先检测人再定位点（准、慢）/ 先找所有点再组人（快、挤）；两阶段混合是目前主流。
遮挡	occlusion	身体部位被自己、他人或物体挡住；是姿态估计最核心的困难，也是"高置信度错误"的主要来源。
抖动	jitter	相邻帧间关键点位置的随机跳动；成因包括噪声、遮挡、模型敏感性与运动过快。
时序平滑	temporal smoothing	利用相邻帧信息抑制抖动；常用移动平均、指数移动平均、巴特沃斯滤波、卡尔曼滤波。
归一化抖动指标	normalized jitter metric	相邻帧关键点位移中位数 ÷ 躯干长度；只用双方都过阈值且相邻的帧对计算，用于比较平滑前后。
骨架图结构	skeleton graph	把关节当节点、骨骼当边的图结构；时序上堆叠后成为时空图，可输入图卷积网络。
姿态条件	pose conditioning	把骨架图或姿态数据作为生成模型的约束条件，锁定关节位置与肢体朝向。
人体参数化模型	SMPL / SMPL-X	用"姿势参数 + 体型参数"描述人体的统计模型；把关键点"装"进一个完整的人体。SMPL-X 还含面部与手。
单目三维歧义	monocular depth ambiguity	单张图像无法唯一确定深度——同样的二维投影对应无数种三维姿势。
重定向	motion retargeting	把一段动作从源骨架搬到目标骨架；三道关：骨架映射、比例适配、足部锁定。
足部锁定	foot lock	把支撑脚钉在地面上，消除滑步；按动作语义开关（接触地面开、腾空关）。
比例适配	proportion adaptation	源角色与目标角色骨骼长度不同，需把关节旋转"翻译"到目标骨架，并处理由此产生的偏移。
滑步 / 穿模 / 镜像翻转	foot skating / clipping / mirror flip	三类典型失败：脚下打滑、肢体穿透、左右互换。处方都是"改输入优先于调参数"。
一致性	consistency	角色跨帧、跨镜头保持"还是同一个人"的能力；本课程中由角色设定卡、参考图与姿态数据共同保障。

6.10 动手实践：从一段自拍到一个角色动作

本章的实践是一条完整链路：**拍摄与提取 → 驱动图像角色 → 驱动三维角色 → 对照实验 → 故障排查**。它同时也是期末作品《春江花月夜》动作段落的产线。先看全景，再逐项动手。

动笔之前提两条组织建议。其一，**按角色分工而不是按步骤轮流**：三人小组可以设"拍摄与素材手"（负责机位、服装、标定自检）、"提取与质检手"（负责置信度过滤、抖动测量、曲线检查）、"生成与记录手"（负责参数调整、结果归档、技术说明素材整理）；每人主负责一环，但要参与全部环节的交叉验收——只做自己那一环的人，永远发现不了"上游的错误在下游长什么样"。其二，**给"重拍"留出预算**：本章案例里平均每个可用片段的背后是约三次生成尝试，动作拍摄同样适用"一条不行再拍一条"的预期，提前和出镜的同学约好额外的时间，比现场硬凑要体面得多。

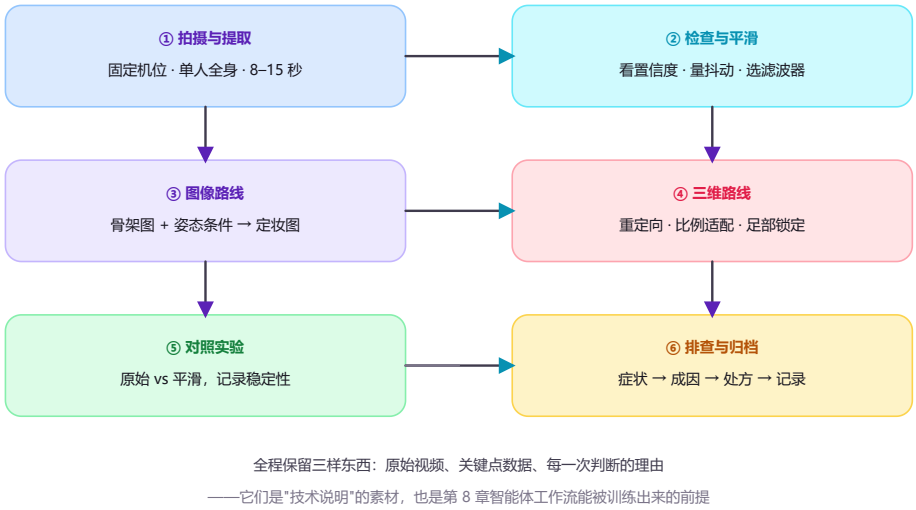


图 6-16　实践链路全景。六步里有两步（②⑤）是本章独有的"质检基因"：**不做质检的动作迁移，等于闭着眼睛生成。**

🛠 实践一：拍一段能被"读懂"的动作（40 分钟 • 课后完成）

基础

找一面白墙或干净背景，用手机三脚架（或把手机架在书堆上）固定机位，横向拍摄自己完成一个完整动作单元——建议用"拂袖回望"或"伸臂转身"，全程 8–15 秒，动作从中速开始。

拍摄前自查（对照表 6-14）：① 全身入框（头顶与脚下各留一掌空隙）；② 单人，画面里没有其他人；③ 光线均匀，身上没有大片阴影；④ 穿有清晰肩线、袖口不过肘的衣服；⑤ 镜头与腰同高、与身体垂直。

提取：把视频导入姿态提取工具（网页端演示工具即可，无需安装），导出关键点数据与叠加显示的视频。逐帧检查：手、脚、肘、膝是否稳定跟随；记录出现问题的帧号。

提交物：一段原始视频 + 一张姿态叠加截图 + 一份"问题帧清单"（哪几帧、哪个部位、现象一句话）。

表 6-14　拍摄与采集自查清单

项目	合格标准	不合格时会出现什么
取景	单人全身入框，头顶与脚下各留一掌空隙	脚部截断 → 无法足部锁定；头顶截断 → 头部关键点漂移
机位	三脚架固定，镜头与腰同高、与身体垂直	镜头晃动被误读为角色运动 → 滑步、漂浮
光线	均匀光照，避免强侧光与明暗分界扫过身体	特征一致性下降 → 关键点跳变
服装	清晰肩线、袖口不过肘、衣摆不过膝	宽大袖袍遮挡手肘 → 手部畸变；长衣摆遮挡接触点 → 无法锁定脚
动作节奏	中速，动作之间有停顿与呼吸	过快动作只抽到少数有效帧 → 输出短于输入、动作断裂
时长	单个动作单元 3–30 秒，建议 8–15 秒	过短缺数据锁定动作模式；过长丢主体一致性
标定动作	开拍先做"举手、张手、踩实"各 2 秒	无法快速判断手部与脚部能否被稳定检出
多人	画面中只有一个人	多人 ID 切换 → 关键点在两个人之间"跳"

🛠 实践二：让图像角色跟着动起来（50 分钟 • 课后完成）

基础

第一步，做骨架图。从实践一的结果里选一帧姿势最好的关键帧，导出为骨架图；按表 6-9 的三条规范检查：**画布尺寸与目标出图尺寸一致**（如 512×512）、纯黑背景白色骨线、没有多余装饰。若工具支持手部关键点，勾选"含手"选项。

第二步，姿态条件生成。把骨架图作为姿态条件输入（控制强度从 1.2 起步），角色图/提示词用你自己设定的角色——如果你还没定角色，用本书的诗人形象或任意一个与你期末作品一致的角色。

第三步，做三组对照。只改一个变量：控制强度依次取 1.0、1.3、1.5，其他参数不变。记录三组结果中"手部是否变形""姿势是否贴合骨架图"的差异。

提交物：三组对照图 + 一段 100 字以上的结论（你的角色适合的控制强度区间是多少，依据是什么）。

✎ 提示词　　姿态条件生成的提示词写法

主体　明确角色：古装诗人、白色长袍、束发　动作　写出动作短语：`raise sleeve, turn head to look back`　场景　江畔、薄雾、月夜；与短片其他镜头一致　负向　`deformed hands, extra fingers, bad anatomy, blurry`　**关键**　姿态由骨架图锁定，提示词不要重复描述姿势细节——让条件去管姿势，让文字去管"这是什么、在哪里、什么风格"

🛠 实践三：把动作交给第 5 章的三维角色（60 分钟 • 课后完成）

进阶

准备：打开你在第 5 章完成绑定（骨骼与蒙皮）的诗人角色，以及实践一提取的动作数据（导出为 BVH 或 FBX 格式的动作文件）。

第一步，骨架映射。检查两套骨骼的对应关系：源（姿态数据）的髋、肩、肘、腕、膝、踝，能否一一映射到目标角色的对应骨骼。参考第 5 章的骨骼命名规范——**命名对得上，映射就快**；对不上时逐根手动画对应线。

第二步，比例适配。播放一次，重点看三处：手臂伸展时手是否戳进躯干、弯腰时头是否没入胸口、迈步时腿的弯折是否正常。发现明显别扭时，调整重定向的比例映射参数。

第三步，足部锁定。找出所有"应该踩地"的时间段（支撑脚），开启足部锁定；确认抬脚-落脚的瞬间没有打滑。如果你的动作含跳跃或滑行，请把那些时段排除在锁定之外。

提交物：一段 10 秒左右的三维角色动作预览（屏幕录制即可）+ 一段说明（你改了哪些参数、为什么）。

🛠 实践四：对照实验——原始关键点 vs 平滑后（60 分钟 • 课后完成）

进阶

目的：这是本章指定的对照实验。用同一段动作数据，验证"时序平滑"到底改变了什么——不是看谁更"好看"，而是用一个可复现的数字说明差别。

做法：① 对实践一的关键点数据，先做置信度过滤（低于阈值 0.5 的点先标记为缺失）；② 取一份**不做任何平滑**的数据作为对照组；③ 取一份用指数移动平均（EMA）或巴特沃斯滤波处理过的数据作为实验组（滤波器按表 6-6 选型）；④ 对两组分别计算**归一化抖动指标**：相邻帧关键点位移的中位数 ÷ 躯干长度，且只用"两帧都过阈值、且帧号相邻"的帧对；⑤ 把两组的结果填进表 6-15。

观察三件事：① 平滑后抖动指标下降了多少？② 有没有"过度平滑"的代价——快速挥手、跺脚等瞬时动作变得绵软、幅度缩水？③ 把手部单独拎出来算：手部的抖动通常比躯干大得多，平滑对它的改善是否更明显？

提交物：填好的表 6-15 + 一条曲线或柱状对比图 + 一段结论：你这套参数用在你的拍摄素材上，是不是"更平滑 = 更好"？请给出证据。

表 6-15　对照实验记录表（原始关键点 vs 平滑后）

测量项	对照组：未平滑	实验组：平滑后	变化
低置信度点比例			
缺失点帧比例			
整体归一化抖动（全部点）			
手腕等末梢点归一化抖动			
快速动作段幅度保留情况			
主观观感（三人独立打分）			

对照组：未平滑

输出保留了动作的全部细节与瞬时峰值，快速挥手时幅度真实。

代价：末梢点抖动明显，手腕与脚踝在每帧间小幅"颤动"；迁移到角色上时，角色的手指与衣摆跟着颤。

适用：需要保留爆发力与写实感的短镜头，或平滑会导致动作绵软的场景。

实验组：EMA 平滑

抖动指标显著下降，末梢点明显稳定，动作整体更"顺"。

代价：引入可控延迟，快速动作峰值被削——挥手到位的瞬间幅度略减、回到中位的速度略慢。

适用：慢表演、舞蹈段落、需要"干净"输出的成片镜头。

🛠 实践五：故障排查演练（30 分钟 • 课堂讨论）

挑战

玩法：三人一组，每人从本章失败诊所中选一个症状（或由教师提供故障样本截图/视频），另外两人**只看症状、不看答案**，用表 6-16 的排查清单，从"症状"推到"成因"再开出"处方"。

规则：① 每次只允许改一个输入或一个参数；② 每改一项都要说出"我预期会发生什么变化"；③ 结果与预期不符时，记录这条"反例"——反例比成功更值钱。

提交物：一张排查记录（症状 / 尝试 / 预期 / 实际 / 结论）。这一项也纳入期末技术说明的素材。

表 6-16　动作迁移排查清单（按症状）

看到什么	先查什么（按顺序）	最常见的真凶
脚在打滑	足部锁定是否按语义开启 → 参考视频机位是否稳定 → 支撑脚是否被遮挡 → 比例适配参数	未做足部锁定 / 机位晃动
肢体穿透身体	是否开启碰撞 → 是否设计了并发动作 → 动作幅度参数 → 极端动作是否超限	并发动作 / 幅度过高
手指粘连多指	参考手部是否被遮挡 → 提取器是否含手部点 → 负向提示词 → 是否做了只修手的局部重绘	手部遮挡引发的提取错误
左右互换	是否单侧被遮挡 → 两侧姿势是否对称 → 是否误开了镜像选项	遮挡 + 对称输入的组合
动作幅度失真	控制强度 / 幅度参数 → 首帧角色图比例 → 参考与角色的尺度是否一致	尺度不一致 / 幅度参数过高
动作整体不像	参考时长是否在 3–30 秒 → 模式是否为高品质 → 是否中途换过模型 → 两个输入源是否冲突	两输入源冲突 / 低质模式
动作忽快忽慢、断裂	是否有丢帧 → 平滑是否过度 → 是否跨模型拼接	丢帧未修复 / 过度平滑

💡 实践总提示 • 三句话

① **先把输入做干净，再开始调参数**——拍摄规范省下的时间，会在生成阶段翻倍还给你；② **每次只改一个变量**——这是把"试出来的"变成"说得清"的唯一方法；③ **把每一次判断的理由写下来**——期末技术说明可以少些形容词，但不能少"我为什么这么选"。

6.11 课程思政：身体影像的使用边界与动作文化

这一章有一个区别于全书其他章节的特殊性：**我们处理的数据是"人"本身。**前几章处理的是文字、图像、视频，这一章处理的是活生生的人体——他的动作、姿态、身体影像。这意味着两件事：第一，**采集之前必须获得同意**；第二，**生成之后可能触犯的法律比别的章节更多、更新。**这一节不打算给你一堆吓人的条文，而是建立一条从拍到发的思考路径。

图 6-17　身体影像使用的三道闸门。从拍到发，每一步都有对应的法律坐标。**把它们背下来不难，难的是在"就拍一下应该没事吧"的瞬间想起它们。**

◆ 一、愿意出镜 ≠ 愿意被"换身"：知情同意要说清什么

一次规范的授权，至少要讲清四件事：**拍什么**（这段动作影像会被提取成姿态数据）、**用在哪儿**（课堂作业？发布到某个平台？参评比赛？）、**改到什么程度**（会被迁移到另一个角色形象上，可能出现在古装人物、卡通形象甚至非人类角色上）、**给谁看**（仅课堂、校内、公开网络）。只说一句"我拍你跳舞"是不够的——同意必须覆盖"将身体影像用于 AI 动作迁移"这件事本身。这不是过度谨慎：动作影像虽然是身体的外在表现，但当它足以指向某个人时，就已经进入了肖像权的保护范围。

为什么"足以指向某个人"这个标准这么重要？因为司法实践中反复确立的正是"**可识别性**"标准：判断是否侵犯肖像权，不看是否直接使用原图，而看能否通过生成的图像识别出特定自然人。北京市互联网法院审理的"AI 换脸"纠纷就是一例：一位古风汉服博主的视频被某 App 换脸模版使用，法院认定不构成"以营利为目的使用肖像"的抗辩不能成立、换脸后的形象仍具有**公众可识别性**，构成对肖像权的侵害。这个逻辑直接延伸到本章：**把同学的动作数据迁移到一个酷似他本人的角色上，即使画面里没有他的原始影像，仍可能构成侵权。**

还有一个容易被忽略的边界：**逝者的身体影像。**民法典第 994 条规定，如无特别规定，死者的姓名、肖像、名誉、荣誉、隐私、遗体等受到侵害的，其近亲属有权请求行为人承担民事责任。用 AI "复活"一位已故者的形象并让其做出他生前未做过的动作，正是典型的越界——近亲属可以主张民事责任。

◆ 二、从"换脸"到"换身"：监管为什么专门盯着生物特征

你可能会问：既然肖像权已经管得住了，为什么还要专门出台深度合成与生成式 AI 的规定？因为这类技术发生的事，**本质上是"伪造一个看起来真实的记录"。**法规要处理的不只是"用了谁的脸"，还有"公众会不会被误导、社会信任会不会被侵蚀"。几条关键规定，请先记住它们的**功能**：

1. **谁采集谁负责。**《个人信息保护法》把**人脸、声纹等生物识别信息**列为敏感个人信息，处理需要**取得单独同意**并履行告知义务（第 28、29、30 条）；本章里"拍摄他人动作并提取姿态"，如果数据能关联到具体个人（例如带人脸的原始视频），就属于这类处理。
2. **工具本身要有防线。**《互联网信息服务深度合成管理规定》要求：提供人脸、声音编辑功能的服务，应当**提示使用者告知被编辑的个人并取得其单独同意**（第 14 条）；提供人脸、声音等人脸生成或编辑功能的模型、模板等工具，须进行**安全评估**（第 15 条）。这一条把责任从"用户"扩展到"平台"——这解释了为什么许多平台会弹窗、会要求你勾选"我已获得肖像权人授权"。
3. **合成要能认出来。**同一规定（第 16、17 条）与《人工智能生成合成内容标识办法》要求对合成内容做**两类标识**：**隐式标识**（写入文件元数据，不改变用户使用体验）与**显式标识**（在合理位置以文字提示呈现，用于可能引起公众混淆的合成声音、人脸生成或替换）。对应到本章的实践：你导出的角色动作视频，应在画面角落或标题中注明"含 AI 生成内容"，且不要把生成工具写入的元数据清理掉。
4. **底线条款。**《生成式人工智能服务管理暂行办法》明确不得利用生成式 AI 侵害他人的肖像权、名誉权、荣誉权、隐私权——这是"红线"而不是"注意事项"。

◆ 三、2026 年最新进展：以"可识别性"为核心的最新规则

本章编写时，知识库里最重要的一条新动态是：**2026 年 9 月 7 日，最高人民法院发布《关于依法审理涉人工智能纠纷案件的意见》（法发〔2026〕10 号）**——这是最高法首部专门针对人工智能纠纷的司法文件，共五个部分二十四条，回应了"AI 换脸拟声""网络开盒""AI 幻觉""大数据杀熟"等社会关切。它没有推翻前面讲的法律，而是把"怎么用"讲得更明确了，其中至少有四处与本章直接相关：

1. **人格权保护的三层结构。**未经同意，用 AI 处理自然人的姓名、肖像等生成**可识别的虚拟数字形象**并使用、发布，构成侵害姓名权、肖像权；未经同意，以自然人声音为训练语料，模仿其音色、语调、发音风格生成**可识别的合成声音**，构成侵害声音权益；操控虚拟形象或合成声音实施不当行为、作虚假陈述，降低他人社会评价的，构成侵害名誉权。
2. **"AI 复活"划出明确边界。**意见明确，未经授权生成逝者的虚拟数字形象，损害其姓名、肖像、名誉的，近亲属可以依据民法典第 994 条主张民事责任。
3. **禁令可以"停服务"。**意见规定人格权侵害禁令除要求行为人停止侵害外，还可以责令网络服务提供者、人工智能服务提供者采取必要措施——在符合比例原则的前提下，可以延伸到"停止提供服务"。对创作者的含义很直接：你上传的内容一旦侵权，处置速度会比过去更快。
4. **标准是"可识别"而不是"完全一致"。**这是最需要写进创作习惯的一条：判断侵权与否的核心标准是**能否识别出特定自然人**，而不是"是否一比一还原"。**企业不能以"相似度低"作为抗辩。**"我改了脸型""只用了 60% 相似度的形象"这类想法，在司法标准面前站不住。

意见同时规定了生成内容侵权的责任分配：引入了类似"通知—删除"的机制，用户恶意输入提示词导致侵权的，由用户承担责任；针对欺诈、侮辱、诽谤、侵犯公民个人信息、传播有害内容等行为，构成犯罪的依法追究刑事责任。

◆ 四、数字虚拟人：更近的一步监管

本章的主题词里天然包含"数字人"的影子。按编写时的公开信息，2026 年 4 月，国家网信部门就《数字虚拟人信息服务管理办法（征求意见稿）》公开征求意见（**注：截至编写时仍为征求意见阶段，请以正式发布版本为准**）。其要点与本课程直接相关：未经**单独同意**，不得生成可识别到特定自然人的数字虚拟人形象；不得利用虚拟人形象丑化自然人、或编造特定身份人员；提供数字虚拟人服务还需进行**算法备案**并定期开展算法安全评估。

同期还有一则行业自律信息值得记住：2026 年 4 月，中国广播电视社会组织联合会演员委员会发表声明，明确标注"非商用""个人二创""公益分享"等字样**不构成免责事由**——"我没赚钱"不能成为未经授权使用他人身体影像的理由。这条对创作心理的纠正很重要：**合法性的门槛在于"有没有授权"，而不是"有没有获利"**。

◆ 五、动作也是文化：尊重不同民族的身体表达

本章的技术处理的是"动作"，而动作从来不只是动作。一个民族的舞蹈步伐、一个地区的仪式身段、一段戏曲的水袖，都是**数代人身体记忆的沉淀**，有它的语法、语境与情感结构，而不是等待被提取的"免费素材库"。请把下面三句话当作创作准则：

- **先理解，再提取。**拍一个民族舞蹈前，先问它是什么场合跳的、谁有资格跳、哪些动作有禁忌含义。把"抬臂转腕"从祭祀语境里剥离出来、贴上"好运舞"的标签，可能冒犯的不只是舞者，而是一整套文化体系。
- **授权要对人，也要对文化。**请传承人出镜、提取其动作数据、迁移到 AI 角色上，三件事需要三层同意：出镜的同意、数据使用范围的同意、风格与文化语境的呈现方式的同意。案例 6-C 里社团的做法可作参考：先书面授权、再约定呈现方式与用途、最后标注 AI 生成。
- **不做猎奇化的呈现。**把不同民族的身体表达统一"翻译"成主流审美的样子，或刻意突出"异域感"，都是一种文化挪用。姿态迁移技术越强，这种风险越大——"我能做到"与"我应该这样呈现吗"是两个必须分开回答的问题。

回到本章的主线：春江小组拍的"捣衣"与"拂袖"，取自古典诗词的意象。处理这类动作时，他们也做了同样的功课：查证"捣衣"是古代妇女在河边用木杵捶打衣料的劳作，动作沉稳、节奏反复，而不是"舞蹈化的道具表演"；正因如此，成片里"捣衣"段落采用的是低机位、长焦的写实处理，而不是加特效。理解动作的生活语境，往往比技术参数更能决定一个镜头的成色。

> 🇨🇳 **课程思政 · 三条可以立刻执行的习惯**
>
> ① **授权优先：**任何真人素材（同学、家人、路人、传承人）进入工作流之前，先取得书面或可回溯的同意，并写清楚"将用于 AI 姿态提取与角色迁移"；② **标识优先：**所有包含 AI 生成或合成内容的作业对外发布时，一律做显式标识并保留隐式标识；③ **克制优先：**不用真人身体影像去生成可能损害他人人格尊严的内容；对他人的身体与文化表达，保持"先理解、再使用"的态度。**技术会更新，这三条不会。**

◆ 把这一章压缩成七句话

1. **姿态估计是把动作变成坐标：**它不"看懂"人，只在每一帧里定位关键点；17 点是通行标准，33 点多了旋转与可见性，133 点补上了手、脸与脚。
2. **每个点都带着"把握程度"：**置信度的含义随架构而异（热图峰值、分类概率、可见性），用法却一致——过滤低置信点、加权三角化、用检出率与缺失率量化质量。
3. **遮挡与抖动是两大天敌：**遮挡会制造"高置信度的错误"，抖动源于噪声、遮挡、快动作与光线变化；归一化抖动指标（位移中位数 ÷ 躯干长）让"有多抖"可测量、可比较。
4. **滤波器都要付出代价：**移动平均延迟大、指数移动平均更平滑、巴特沃斯通带平坦、卡尔曼可融合预测——"更平滑"不等于"更好"，要看动作语义。
5. **三维不是"再加一个坐标"：**单目深度歧义意味着一个投影对应无数姿势；SMPL 类模型用姿势与体型参数描述人体，帮我们把散点"装"进一个完整的人。
6. **重定向过三关：**骨架映射（名字对得上，迁移才快）、比例适配（不同的身长要重新翻译旋转）、足部锁定（接触地面开、腾空关）——滑步、穿模、比例失真都能在这三关里找到出处。
7. **这是一条"数据流"而不是"魔法"：**拍摄规范、置信度过滤、时序平滑、姿态条件（ControlNet-OpenPose 及网页版"动作参考"）、重定向参数——每一步都有输入、有输出、有代价，也都留下了可写进报告的记录；而全程之上的，是"先授权、再标识、后发布"的底线。

❓ 自测 1：COCO 17 点体系里没有手指尖，为什么很多"动作迁移"任务仍然强调使用含手部关键点的工具？

答案：因为"迁移的忠实度取决于条件的密度"。17 点体系只约束到手腕，手掌与手指的位置完全交给生成模型"脑补"，而手恰恰是生成模型最容易出错的地方（指节粘连、多指）。含手部 21 点的体系（如 COCO-WholeBody 133 点中的手部部分、MediaPipe 的手部模型）为生成模型提供了更细的约束，因此手部更稳。但要注意两点：密度更高也意味着更容易被遮挡污染——手部点一旦出错，错误也会被"忠实"地迁移过去；以及并非所有工具都支持手部条件，此时"用袖子挡住手"这种设计层面的规避反而是最省成本的做法。

❓ 自测 2：一位同学把整段民谣歌手的视频丢进工具，提取出全身姿态后（并保留了原始视频），想把动作迁移到角色上。请指出这个流程里至少两处需要先处理的问题。

答案：① **授权问题**：视频里的人是真人（公众人物），把其身体影像用于 AI 姿态提取与角色迁移，需要该人（或其权利人）的同意——公众人物不例外；若原始视频来自网络，还涉及版权问题。② **数据性质问题**：原始视频含人脸，属于生物识别信息（敏感个人信息），处理需单独同意并做"最小必要"控制；只保留关键点数据、删除含人脸的原视频是个常见的降风险做法。③ 还有两个技术层面的坑可加分：整段视频太长（超出 3–30 秒的有效区间）、多人画面会引发 ID 切换。请记住：技术可行与法律可行是两条独立的线，必须分别通过。

❓ 自测 3：解释"置信度 0.9 的关键点也可能是错的"，并说明这对你设计质检流程有什么影响。

答案：置信度衡量的是"模型对自己的判断有多确定"，而不是"这个判断对不对"。模型见过大量人体结构，当关键点被遮挡时会依据先验"猜"一个解剖上合理的位置——它猜得很自信，于是给出高置信度。这就是"高置信度的错误"（如被袖子遮住的手腕）。对质检的影响：① 不能只看置信度，还要看**时序一致性**——单一帧的可信度低，但前后帧突变是可检测的信号；② 对遮挡高发部位（手、脚、被衣物覆盖的关节）单独设检查规则，例如与前后帧的位移超过阈值即报警；③ 拍摄时主动减少遮挡（紧袖口、避开身体自遮挡的角度），是从源头降低这类错误的最经济手段。

❓ 自测 4：滑步（foot skating）在三维重定向和"参考动作驱动角色"两条路线上都会出现。请分别说出各自最直接的一个成因与对策。

答案：三维重定向路线：最直接的成因是源骨架的关节旋转被"翻译"到不同腿长的目标骨架后，脚的世界坐标随之漂移，且没有任何机制把它拉回地面；对策是**足部锁定**——把支撑脚在接触阶段钉在地面坐标上，并按动作语义开关（跳跃/滑行关闭）。"参考动作驱动角色"路线：最直接的成因是模型没有物理模拟，脚与地面的关系是"画"出来的；对策首先在输入侧——固定机位、单人全身、尺度一致、让脚部接触点清晰可见，必要时使用平台提供的足部锁定选项。两者的共同点是：都不是靠"调大某个参数"解决的。

❓ 自测 5：小组想用一位已故诗人的形象，让他"做出"一段现代舞蹈动作，用于校园文化节。请依据本章的思政内容，指出其中的风险与可行的替代方案。

答案：风险有三层：① **逝者人格利益**——民法典第 994 条保护逝者的姓名、肖像、名誉等，近亲属有权主张民事责任；让逝者"做出"其生前未做过的动作，可能构成对其形象的歪曲与不当使用。② **虚拟数字人监管方向**——不得未经单独同意生成可识别到特定自然人的数字虚拟人形象，且不得用于丑化或编造特定身份人员（按征求意见稿的精神），校园文化节这类公开活动无法用"非商用"免责。③ **文化语境**——把历史人物的身体表达与现代舞蹈拼贴，容易造成文化误读。可行替代：用"原创虚构角色 + 古典动作风格"的方式致敬（如本章的诗人角色本就是原创设定）；或在获得逝者近亲属书面同意的前提下，仅做符合其历史形象的静态/克制呈现，并明确标识 AI 生成。核心原则：**技术能"复活"一个人，但决定"该不该"的永远是人。**

❓ 自测 6：为什么本书反复说"改输入优先于调参数"？请用本章的一个具体机制说明。

答案：因为大多数伪影的成因在输入侧，参数只是"放大器"。以滑步为例：若参考视频机位晃动、脚被长衣摆遮住，调任何运动参数都无法恢复"脚接触地面"这一被破坏的信息；相反，改成固定机位、穿露出脚踝的服装（输入侧两项改进），滑步的根源被消除，锁定选项才真正有效。同理，手部畸变的根源是遮挡与条件密度不足，负向提示词只能压制症状；改善拍摄与换取条件才是治本。这也解释了为什么本章的实践设计里，拍摄规范的"位置"与生成参数相当：参数决定"表现上限"，输入决定"错误下限"——而通识课程作品的成色，绝大多数时候由下限决定。

◆ 拓展阅读

- **随查手册：**姿态/动作类工具的选型与额度见附录 B 工具速查；涉及真人身体影像的授权书模板、"发布前三查"与 AI 标识操作，见附录 E 合规清单——本章的合规要求全部落在那份清单里。
- **姿态估计入门：**在任意姿态提取网页工具（如 MediaPipe 官方演示页）上传自己的一段动作视频，观察关键点叠加效果与本底置信度，再看它失败在哪里——亲眼看一次"遮挡误判"，比读十页论文都有效。
- **运动数据资源：**CMU Motion Capture Database——约 2600 条动作捕捉试次，公开可下载（可用于教学与作品，含商用，但不得转售，需保留致谢）；Mixamo——约 2500 条免费动作，适合快速上手三维动画（使用前请阅读其许可条款）。
- **法规原文：**《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》，以及最高人民法院《关于依法审理涉人工智能纠纷案件的意见》（法发〔2026〕10 号，2026 年 9 月）——建议精读与"人格权""标识"相关条款，原文比转述可靠。
- **司法与行业动态：**检索"AI 换脸 肖像权"相关的法院判例（如北京互联网法院的公开案例）与"数字虚拟人 监管"的行业报道，体会"可识别性"标准在真实纠纷中如何被适用。
- **文化延伸：**读一读《春江花月夜》里"捣衣"意象的历代注解，以及民族舞蹈、戏曲身段的科普资料——理解动作背后的文化语义，是让技术作品"有分量"的捷径。

第 7 章

游戏场景生成与编辑

从一句需求到一个能走进去的世界：构成要素、规格表、程序化布局与引擎落地。

学习目标

1. **知识：**理解场景的五个构成要素（地形、建筑、道具、光照、氛围）与三个可用性评价维度（视觉统一、玩法动线、性能预算）；理解程序化生成（噪声地形、波函数坍缩、泊松盘布局）的直觉原理。
2. **能力：**能用大语言模型把自然语言需求转写成一份可校验的场景规格表（JSON）；能按管线生成天空盒、地形、建筑与道具资产；能把资产导入引擎并完成比例、轴向、Pivot、面数的导出前检查。
3. **素质：**建立"结构化需求先于生成"的工程习惯；理解场景设计中的文化准确性与文化挪用问题，尊重考据来源与团队劳动分工。

对应第 9 周 · 3 学时　讲授 2 学时 · 实践 1 学时　主线推进：夜江场景 + 结构化资产清单

> 小舟第一次打开引擎，把昨天生成好的那座"江边楼台"拖进场景视图——结果它像一块巨型广告牌横躺在半空中，比远山还大十倍，还是侧着倒下去的。他反复缩放、旋转、拖动，花了四十分钟才让它"勉强站直"，而旁边的扁舟只有指甲盖大小。小月看着屏幕说了一句本章要回答的全部问题：**"我们是不是从一开始就没说清楚，这个场景到底长什么样？"**

这句话点破了一个常见误会：很多人以为"场景生成"就是"让 AI 画一个背景"。不是的。一个能走进去、能跑的**场景**（scene），本质是一份被严格约定过的工程文件——它规定了有什么东西、放在哪、多大、用什么材质、站在哪儿能看到什么、跑起来要花多少性能。本章要回答三个问题：**第一，一个场景由哪些要素构成，我们凭什么说它"好用"？第二，怎么把一句自然语言需求，变成机器和人都看得懂的规格表？第三，这些资产怎么真正落进引擎，而不是落进"打不开的文件夹"？**顺着这三问，小舟会为《春江花月夜》短片搭出"夜江场景"的初稿——这也是本章的作业。

📌 本章的四族技术：一条流水线上的四种工种

本章不看形状、不看动作，而是看「一个能走进去的空间」。它由**四族技术接力**完成，各管一段路：

- ① **大语言模型**（7.3）= **结构化**：把一句话需求翻译成场景规格表与资产清单
- ② **程序化生成**（7.4）= **算法**：用噪声、规则与采样「摆」出地形与物体散布
- ③ **AI 资产生成**（7.5）= **生成**：天空盒、建筑、道具、材质（复用第 4 章的能力）
- ④ **引擎工程**（7.6）= **落地**：导入、比例与轴向检查、性能预算

前三族的任务是「把它造出来」，第四族是「让它能用」。多数翻车不在前三族，而在第四族——本章失败诊所里最贵的那一堂课（111 倍面数），就发生在这里。

7.1 场景的五块拼图：地形、建筑、道具、光照、氛围

先建立一个总印象。**把场景想成一台话剧的舞台**：演员走的地板是**地形**，两侧的布景和后墙是**建筑**，台上摆的桌椅茶杯是**道具**，头顶那几盏灯是**光照**，而观众感到的"闷热""清冷""雨夜感"是**氛围**。少了任何一块，舞台都会露怯：地板缺了演员就悬空，灯缺了所有人都像纸片，氛围缺了整个画面会像没上色的线稿。

为什么要专门"拆成五块"来讲？因为 **AI 工具是按资产类型分工的。**天空盒生成器只擅长生成 360° 的环境图；三维生成器只知道"我拿到了一个物体的描述"；材质工具输出的是贴图，不是几何。你不先把需求拆成"我到底要哪几类资产"，就会拿着一个工具去要它给不出的东西——这是初学者最常见的浪费。第 1 章讲过的"提示词四要素"里，**输出格式**是最容易被忽略的一环；在场景任务里，"输出格式"就具体成了"输出的是不是这一类资产"。

图 7-1　场景的五块拼图与三层评价。上排是构成要素，中排是它们的叠放顺序，下排是验收标准。"先拆要素、再谈好坏"，是本章全部内容的总纲。

下面逐块说明，并且每一块都给出"A 类工具缺口"——也就是 AI 最不擅长的那一小段，这正是你未来最值钱的手艺所在。

◆ 地形：场景的"地基"

地形（terrain）负责回答"脚踩在哪里、视线在哪里被挡住"。它包含两部分：**高度场**（地面起伏，通常用一张灰度图表示"每个点多高"）和**地表材质**（草地、沙石、泥土的分布）。水面比较特殊，它既像地形（有高度、反射天空）又像道具（有波纹、有倒影），工程上一般单独处理，本章在 7.5 节专门讲。

生活类比：地形就是**房子的地面平整度**——地不平，家具怎么摆都别扭。课程情境：小舟的"夜江"里，地形不能是"一座山加一片平地"，而要有一条从远山切到江岸的缓坡，让镜头从岸边推向江心时有自然的高差。**AI 工具缺口：**现成的文生三维工具只生成"一个物体"，不生成"一片几平方公里的地形"；地形必须靠程序化算法（7.4 节）或地形编辑工具生成。

◆ 建筑：画面的"骨架"

建筑（architecture）承担两个职责：一是**体量**——在远景里提供剪影和层次，让画面有"远近"；二是**形制**——在近景里提供文化信息，让观众一眼认出"这是哪里的、哪个年代的"。第二点在后文思政部分会重点展开，因为它最容易出错。

生活类比：建筑像**一幅画的构图骨架**，它先于细节存在。课程情境：小江希望"楼台"出现在画面右三分之一处、半掩在柳树后——这不是美术偏好，而是**规格**：位置 (x, y, z)、朝向角、高度、占地范围，都可以写成数字写进清单里。

◆ 道具：故事的"标点符号"

道具（props）是场景里那些"可以被拿走、被使用、被特写"的小物件：一盏灯、一条船、一张琴、一摞书。它们体量小，但对叙事极其关键——**道具是导演用来提示观众的标点**：逗号、顿号、感叹号。

课程情境：《春江花月夜》里的"扁舟"就是核心道具。它小，但它是"人"的替身——船停在水上，观众就会自动补上"有人在这里看月亮"。这类"小而不重要"的资产，恰恰是 AI 生成器最擅长、单价最低的一类，也是最适合练手的起点。

◆ 光照：决定"像不像"的一半

光照（lighting）包含三个可调的东西：**光源**（太阳、月亮、灯）、**方向与强度**、以及**全局光照设置**（环境光从哪里来、阴影怎么衰减）。一个被反复验证的经验是：**同一个模型，换一套光照，观感差距可以大于换一个模型。**这也是为什么很多看起来"很高级"的场景，其实模型面数并不高。

课程情境：夜江场景的主光源只能是月亮。这意味着光的方向要统一（所有景物的受光面朝同一侧）、强度要压得低、色温要偏冷，而楼台窗内的一点暖光成为画面里唯一的暖色——**这个"冷暖对撞"就是全片的气氛支点。**

◆ 氛围：最便宜、也最容易做砸的一层

氛围（atmosphere）是天空、雾、色调映射与粒子（水汽、飞萤、落花）的总和。它最便宜——一个天空盒加一层雾就能让画面"换季"；也最容易做砸——因为它是**全局层**：改一次雾的浓度，所有资产的颜色都要重新对一遍。

生活类比：氛围像**给照片加的那层统一滤镜**；滤镜能让同一个房间 morning 和 midnight 判若两地。课程情境：如果天空盒是冷蓝、雾是中性灰，而楼台却是暖黄、扁舟是青绿，那台上所有东西都在"各自发光"，观众的第一反应是"假"。所以本书建议一条硬规矩：**先定氛围，再生成资产**，并把色板写进清单。

表 7-1　场景五要素：职责、典型资产与出错信号

要素	回答什么问题	典型资产	AI 可用的手段	出错信号
地形	脚踩在哪里、视线被什么挡住	高度场、地表材质、河床	程序化噪声生成 / 地形编辑工具雕刻	地面像塑料布一样平；坡度走不上去
建筑	剪影与层次从哪里来、这是哪里	楼台、院墙、桥梁、牌坊	文生三维 / 图生三维；参考图 + 多视角质检	形制混搭、比例失真、悬空或陷地
道具	故事用什么标点提示观众	扁舟、灯、琴、书案、樽	文生三维（小物件成功率最高）	数量太多、满地散落、没有叙事重点
光照	光从哪来、明暗关系对不对	方向光、点光、环境光探针	引擎内手调 + 天空盒提供环境光	多处高光冲突；夜景却有白天的亮度
氛围	这是什么时候、什么天气、什么心情	天空盒、雾、色调映射、粒子	文生天空盒（360° 等距柱状图）	天空与地面冷暖打架；雾把远景糊死

∫ 深入一格 · 场景其实是一棵树：层级与变换 选学 · 数学

为什么"看起来很小的一步操作"会在场景里造成灾难？答案藏在场景的组织方式里：引擎里的场景不是一张平面的清单，而是一棵**树**（层级，hierarchy）。每个物体是一个节点，节点之间可以建立父子关系：把"灯笼"挂在"船"下面，灯笼就成了船的**子节点**。

父子关系带来一个极强的性质：**子节点的最终位置 = 父节点变换 × 子节点自身变换**。用矩阵写出来就是：

$$世界变换 = 父级变换 \times 局部变换$$

父级移动 1 米，所有子节点跟着走 1 米；父级旋转 90°，所有子节点的朝向一起转——这就是"拖一下动一片"的原因

这个性质一半是祝福，一半是诅咒。**祝福**：做一艘会动的船，只要移动"船"这个父节点，船上的桨、灯、篷全都会跟着动，不用逐个改。**诅咒**：如果你把整个场景的资产都堆在一个父节点下，而这个父节点自己带着一个"未应用"的缩放（比如 0.01），那么屏幕上一切正常，导出后所有资产都会变成原来的百分之一——这正是小舟遇到问题的技术根源之一。

还有一个与它相关、但更难看见的概念：**Pivot**（轴心点），也就是"这个物体的原点在哪、旋转围绕哪个点转"。Pivot 不显示在画面上，却决定了放置的手感：一扇门的 Pivot 应该在门轴上，如果它在模型的几何中心，那么你旋转它时门会绕着"腰"转，而不是绕着"合页"转——这类问题在渲染截图里完全看不出来，只有在场景里操作时才会暴露。7.6 节会给出一份检查表专门处理它们。

🎯 一句话记住五要素的分工

地形给"地"，建筑给"形"，道具给"事"，光照给"真"，氛围给"情"。做场景卡壳时，先问自己：**我现在缺的是哪一块？**——大多数"看起来不对"的场景，问题出在光照和氛围，而不是模型本身。

7.2 场景好不好用：三个评价维度

"好看"是一个无法验收的词。工程上必须把它拆成可检查的指标，才谈得上改进。场景的可用性可以拆成三个维度，它们互相拉扯，构成一个经典的"不可能三角"：**视觉统一、玩法动线、性能预算——你很难同时把三个都拉满**，成熟的做法是"明确谁优先、谁让步"。

◆ 维度一：视觉统一（风格一致 + 构图合理）

视觉统一说的是：**场景里的所有东西，看起来是不是出自同一个世界**。它包含内容与形式两层——风格（写实/水墨/低多边形/像素）、比例（人的身高与门的高度是否匹配）、色彩（是否共用一套色板）、细节密度（远处的石头和近处的石头精细程度是否合理）。

要诚实说明一件事：**截至本书编写时，检索范围内并没有见到针对"AI 场景风格统一性"的公认量化标准**。行业里更常见的做法是用工程手段替代标准：例如天空盒生成工具 Skybox AI 提供 30 余种**风格包**（Style Pack）预设，让同一套场景的天空、光照保持一致；三维生成侧的 Seed3D 2.0 则通过**统一的 PBR 生成模型**来减少多步骤之间的"级联误差"（每一步都偏一点，最后整体跑偏）。

对课程作业，本书给出一套可操作的"土办法"，比追求量化标准更管用：**一份参考图集（3–5 张）+ 一个色板文件 + 一套材质命名规范**。所有资产生成前先过一遍参考图集，生成后对照色板看一眼——统一性靠"减少随机"，不靠"事后调色"。

◆ 维度二：玩法动线（可通行、看得懂、走得通）

场景不只是被看的，还是被"走"的。即使你的作品是短片而不是游戏，动线（玩家/镜头在空间中的移动路径）也决定观众能不能看懂。动线包含三件事：**起点与终点**（从哪进来、到哪结束）、**可通行性**（哪里能走、哪里是墙）、**引导与遮蔽**（用什么吸引注意力、用什么挡住不该看的东西，行业里叫"掩体/遮挡"）。

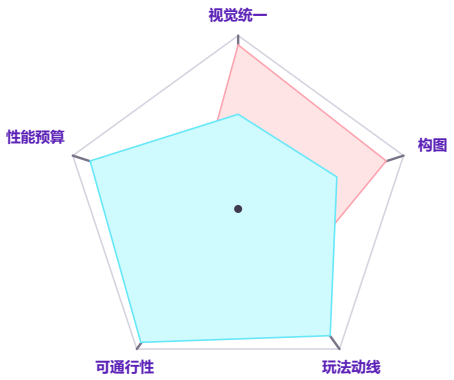
这里有一条来自真实项目的宝贵经验：**校验必须先于实例化**。检索到的多个 AI 关卡生成项目都遵循同一个流程——先让模型生成布局数据，**对布局做严格校验**（确保从出生点到钥匙、门、出口确实存在一条通路），校验通过之后才真正把物体实例化进引擎。其中 WATCH MY ESCAPE 的工作流更进一步，坚持**先设计解法路径、再写地图 JSON**：先想清楚"怎么算通关"，再倒推地图长什么样。

课程情境：小舟把这条经验移植到短片里——先定 6 个镜头的机位与运动路径，再让楼台、扁舟、柳树的**位置服务于这些机位**（哪几个物体必须在画面里、哪个物体必须挡住画面外的东西）。这样搭出来的场景是"给镜头用的"，而不是"摆着好看的"。

◆ 维度三：性能预算（面数、贴图、LOD、绘制调用）

性能预算（performance budget）是三个维度里最容易被忽略、却最"硬"的一个：它像**一个月的伙食费**，花超了就要挨饿。场景的性能开销主要来自四笔账：**三角面总数**（模型精细度）、**贴图内存**（每张 2K/4K 贴图都要占用显存）、**细节层级（LOD）**是否分层、以及**绘制调用（draw call）**的次数（引擎每提交一次"画这个材质的一批物体"就是一次调用）。

关于绘制调用，本书要坦白一个"事实底座"里的空白：**检索范围内没有拿到 2026 年的 draw call 预算标准数字**，所以这里只讲原理不给具体配额——**合并材质、减少材质数量、对重复物体做实例化，是降低绘制调用的三个基本手段**。同理，具体的面数预算也要看目标平台与引擎版本，本书在 7.6 节给出一份行业常用的参考区间，并明确标注它的适用条件。



方案 A • "好看但跑不动"
视觉与构图拉满：全高模、无 LOD、每件资产独立材质
代价：面数失控、绘制调用暴涨，实时预览掉帧
典型来源：只按"生成得漂亮"挑选资产
适用：单帧渲染的静帧作品、离线出图

方案 B • "能跑但平淡"
动线、通行与性能合格：低模 + LOD + 共享材质
代价：远景单薄、细节密度不足，画面缺"重量"
典型来源：直接使用生成器的低面数默认档
适用：交互式漫游、需要稳定帧率的游戏原型

没有"三个顶点"的方案——先决定谁优先，再让谁让步。
本课程作业建议：性能优先（保证能演示）→ 动线其次 → 视觉统一最后用统一色板与光照"补"回来。

图 7-2　三个维度互相拉扯：两种极端方案。雷达图越靠外越好。多数失败场景不是"处处差"，而是像方案 A 那样——在一个维度上过度投入，在另一个维度上彻底失控。

表 7-2　三个维度的可操作检查项

维度	子项	怎么检查（不依赖主观感受）	不通过时的第一处方
视觉统一	风格一致	把全部资产放进同一张对照图，遮住标签后能否一眼认出"这是一套"	统一参考图集与色板，重生成跑偏的资产
	比例关系	放入一个参照物体（如 1.8 米高的人形盒），逐个比对高度	按参照物缩放并应用变换，再重新导出
玩法动线	可通行性	用胶囊体沿规划路径走一遍，记录卡住的位置	调整坡度、清除阻挡、加宽通道
	视线引导	在关键机位截图，圈出画面第一眼落在哪里	用光照与材质对比度强化主体
性能预算	面数与贴图	查看场景统计面板的三角面总数与贴图内存	减面 / 重拓扑 / 压缩贴图
	LOD 与材质	检查远景资产是否有低模替代；统计独立材质数量	生成 LOD 链；合并材质

> 📌 **路线图：本章接下来怎么走**
>
> 有了"五要素 + 三维度"这把尺子，后面四节就是一条流水线：**7.3 把需求变成规格表**（用大语言模型）→ **7.4 用程序化算法铺出地形与布局**（噪声、WFC、泊松盘）→ **7.5 按资产类型生成内容**（天空盒、建筑、道具、材质）→ **7.6 落进引擎并把性能账算清。**7.7 之后是案例、失败诊所、动手实践与思政专节。

7.3 把一句话变成一张清单：大语言模型与场景规格表

回到小舟的困境。他说的是"搭一个夜江场景"，而引擎需要的是"这 47 件资产，各自的编号、类型、位置、朝向、尺寸、面数上限和材质"。这两者之间隔着的不是"美术水平"，而是**一次翻译**——把自然语言需求翻译成结构化数据。

好消息是：这次翻译恰好是**大语言模型（large language model, LLM）最可靠的一种用法。**坏消息是：它只有在被"约束"时才可靠。本节讲清楚这个约束怎么做。

◆ 核心模式：一句话 → JSON 规格 → 校验 → 建场景

把检索到的多个可运行项目放在一起看，会发现它们不约而同地长成同一个样子：**自然语言描述关卡 → 由模型生成 JSON 格式的场景/布局规格 → 对 JSON 做 schema 校验与游戏规则校验 → 解析进引擎，构建出可玩的场景。**所有项目反复强调的关键词是同一组：**严格的输出契约、schema 校验、解析失败后的重试、强制结构化输出而不是解析自由文本。**

为什么这个顺序不能换？用一个生活类比就懂了：**它像装修。**你不会对施工队说"给我弄个温馨点的客厅"，然后指望他们自己拿主意；你会先出一张**施工图**——哪里是承重墙、插座几个、柜子多宽。施工图就是规格表：它把"温馨"翻译成可以核对的数字。而"先校验再施工"对应装修里的**验线**：如果图纸本身自相矛盾（门开在承重墙上），越早发现越省钱；等到墙都砌完了才发现，代价就是全部重做。

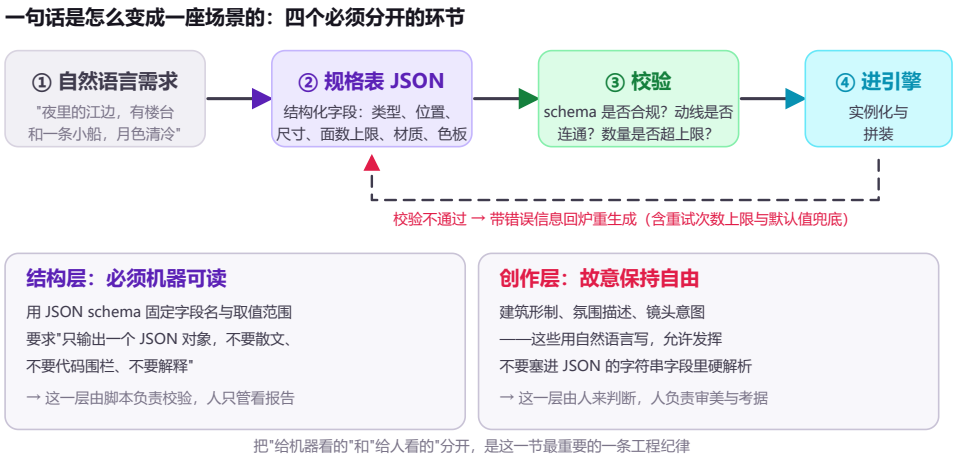


图 7-3　从自然语言到场景的四步流水线。③ 是多数人会跳过、但决定成败的一环；右下角"结构层 / 创作层分离"则是让整条流水线不互相拖累的关键设计。

◆ 六个真实项目，一个共同答案

下面这张表是检索范围内可查到的、把"LLM 生成场景"真正做成了工具的项目。它们的做法差异很大，但收敛到同一批工程要点上——这些要点就是你可以直接抄进自己作业的模板。

表 7-3　六个"LLM 生成场景"项目的做法与可借鉴要点

项目	做法	可借鉴要点
AI PCG - 2D（Unity 资源商店，发布日 2026-06-22）	用具备空间推理能力的模型生成 JSON，描述一个完整、逻辑可通关的 2D 地牢布局；编辑器导入器解析 JSON，并按严格连通性规则校验（确保出生点到钥匙、门、出口存在通路），通过后才实例化到瓦片地图	**校验先于实例化**；提示工程可以被专业化成"主关卡设计师"式的专家系统
AI Level Designer（开源，Unity 6 编辑器扩展）	把自然语言提示经在线模型或本地模型转成 JSON 场景蓝图；模型的回答**必须遵循内部 schema**；内置 schema 校验、目录约束（最大对象上限）、网格与世界坐标模式、以及面向不同游戏类型的模板	**schema 契约 + 硬上限 + 类型模板**三件套；"目录约束"能直接防住"生成 500 棵树"
game-exa 的 game-designer 技能	把自由形式的游戏描述转成严格 schema 的设计文档 JSON，其输出驱动贴图生成、关卡设计与代码合成；输出契约写明：**只输出单个 JSON 对象，无散文、无代码围栏**；并附硬约束（如恰好一个玩家实体）与校验脚本	**"只输出 JSON"的硬契约**；用"恰好一个"这类断言把逻辑错误挡在门外
WATCH MY ESCAPE 的 generate-map 技能	从情景描述生成并校验可玩地图文档：固定 15×15 网格、带行为的实体、谜题逻辑、**分步解法走查**；工作流坚持**先设计解法路径、再写 JSON**	**可解性先于可玩性**；把"解法"当作一等公民先做出来
Grue.is 的提示词指南	用**工具调用**做结构化输出：用规划、内容生成、玩法更新三类函数承载数据，并以强制指定函数的方式产生稳健 JSON；建议往工具 schema 加字段而不是写自由文本，schema 版本化并显式标注必填字段，**在应用状态变更前先校验模型提出的变更**	**用强制函数调用代替解析自由文本**；把"接受模型建议"变成一个需要校验的动作
lore-weave	分层方法：结构层用结构化 JSON，**解析失败时重试**（重试耗尽则回落到规范默认值）；叙事层则刻意保持自由散文、不做结构校验；同时跟踪提示模板版本号，schema 变更导致缓存输出失效时提升版本号	**结构层与创作层分离**；**重试 + 默认值兜底**；提示模板版本号
UniKit AI	规范驱动：把游戏设计写成机器可读的注册表文件，规划从中读取；编辑器操作（场景、预制体）**经由引擎的 MCP 服务器执行**；词汇表已覆盖三款主流引擎	**"规格表 + 执行器"的架构**——第 8 章讲智能体时会再遇到 MCP 这个词

学术侧也有两条对应的证据链：其一，一篇 2026 年 2 月发表的会议论文《Hybrid Procedural World Generation Using Generative AI》用一个网页应用把模型产出的**结构良好、清晰的 JSON 数据**（描述地形、任务叙事与物品特性）直接集成进游戏引擎；其二，一篇题为《Knowledge-Conditioned, Single-Pass LLM Synthesis of Executable Unity Game Scenes》的预印本证明了"带 schema 约束"的收益——在**带 schema**的条件下，用一个两段式管线（先产出引擎特定的**中间表示** JSON，提示词明确要求"只输出合法 JSON，不要多余文本"，再由编码提示词从中间表示生成引擎脚本）就能合成可执行的场景，而论文同时编目了若干编译错误类型，提醒我们"能生成"与"能编译"仍是两回事。

> **🎯 五条最佳实践（可直接抄进你的作业）**
>
> 1. **强制结构化输出**——用 JSON schema 或带强制函数选择的工具调用，而不是解析自由文本。
> 2. **建场景前先校验两件事**——schema 合规，以及玩法规则（连通性、可解性、数量上限）。
> 3. **解析失败要能重试**，并用默认值兜底；schema 与提示模板都要版本化，避免旧缓存污染新结果。
> 4. **先设计并校验"解法逻辑"**（机位路径、开门条件），再谈摆件。
> 5. **把创作/叙事层（自由散文）与结构层（严格 JSON）分开。**

◆ 规格表长什么样：字段设计与一份可用的模板

字段设计有三个原则：**机器能校验**（用数字、枚举、布尔值，不用形容词）、**人看得懂**（字段名直白，附单位）、**不重不漏**（每件资产只归属于一个类别，每个相机只引用已存在的资产编号）。下面这张表把字段分类说明。

表 7-4　场景规格表的核心字段（第 7 章作业模板）

分组	字段示例	取值要求	为什么必须有
身份	`scene_id / version / engine`	字符串；版本号用数字递增	让场景可追溯：改了哪一版、给哪个引擎用
单位与范围	`units / extent.x_m / extent.z_m`	单位固定为米；范围用数字	尺度错误的根源往往就是"单位没说"
地形	`generator / octaves / lacunarity / persistence / max_height_m / water_level_m`	枚举 + 数字；水面高度与地形同坐标系	把 7.4 节的噪声参数直接写进清单，结果可复现
氛围	`skybox_prompt / time_of_day / palette / fog_density`	色板给出 3–5 个十六进制色值	色板是"视觉统一"唯一可传递的东西
资产条目	`assets[]: id / type / prompt / position / rotation_deg / size_m / max_tris / material / lod_levels`	id 唯一；type 限定枚举；max_tris 给硬上限	条目是清单的最小单位，也是"目录约束"的载体
预算	`budget.triangles_total / budget.texture_mb`	数字；与资产	7.6 节的性能

分组	字段示例	取值要求	为什么必须有
		条目自洽	检查要靠它给出红线
镜头与动线	`cameras[]: id / position / look_at / path`	look_at 只能引用已存在的资产 id	让场景"为镜头而建"，而不是摆完再说

下面是一份可直接使用、也可以直接复制去改的 JSON 骨架。注意它**通篇只有数字、枚举和短标识符**，没有一句抒情描写——抒情的部分（"月色清冷"）被我们留在了提示词里，因为那是给人看的。

```
{
  "scene_id": "chunjiang_night_river_v1",
  "version": 1,
  "engine": "Unity 2022 LTS",
  "units": "meter",
  "extent": { "x_m": 320, "z_m": 320 },
  "terrain": {
    "generator": "fbm", "seed": 20261009,
    "octaves": 5, "lacunarity": 2.0, "persistence": 0.5,
    "max_height_m": 18, "water_level_m": 0
  },
  "atmosphere": {
    "time_of_day": "night",
    "palette": ["#0b1b33", "#1d3557", "#6d8fae", "#e8d5a3", "#f4a261"],
    "fog_density": 0.008
  },
  "assets": [
    { "id": "sky_001", "type": "skybox", "format": "HDR",
      "prompt": "夜空，满月高悬于江面上方，薄云，远处山影，冷蓝主调" },
    { "id": "terr_001", "type": "terrain", "max_tris": 40000 },
    { "id": "water_001", "type": "water", "level_m": 0,
      "material": "M_Water_Night", "max_tris": 200 },
    { "id": "bld_001", "type": "building", "name": "临江楼台",
      "position": [-6, 0, 22], "rotation_deg": [0, -18, 0],
      "size_m": [9, 7, 6], "max_tris": 8000, "lod_levels": 3 },
    { "id": "prop_001", "type": "prop", "name": "扁舟",
      "position": [4, 0, 30], "rotation_deg": [0, 12, 0],
      "size_m": [3.2, 0.6, 1.1], "max_tris": 1200, "lod_levels": 2 },
    { "id": "prop_002", "type": "prop", "name": "灯笼",
      "position": [-3, 2.4, 23], "size_m": [0.4, 0.5, 0.4], "max_tris": 400 }
  ],
  "budget": { "triangles_total": 120000, "texture_mb": 180 },
  "cameras": [
    { "id": "cam_01", "position": [0, 1.6, 40], "look_at": "prop_001" }
  ]
}
```

✎ 提示词 · 生成场景规格表（可复制）

你是一名资深关卡设计师，负责把文字需求转写成 JSON 场景规格。下面是内部 schema 的字段要求：**[粘贴表 7-4 的字段清单]**。
需求描述：**[夜景江边，一轮明月，右岸有一座临江楼台，水面上有一条扁舟，月色清冷、雾气淡]**。
硬性约束：
1) 只输出一个 JSON 对象，不要散文、不要解释、不要代码围栏；
2) 所有长度单位为米，水面高度与地形使用同一坐标系；
3) 资产条目不超过 **[12]** 条，id 唯一且格式为 **[类型缩写_三位序号]**；
4) 每件资产必须给出 max_tris，建筑不超过 **[8000]**、道具不超过 **[1500]**；
5) 相机的 look_at 只能引用已存在的资产 id。
若某条需求无法映射到字段，请在 JSON 的 notes 字段里用一句中文说明，不要自行发明字段。

输出 JSON　温度 0.2–0.4　模式 推理模式更稳　重试 最多 3 次

这份提示词里有三处值得单独说明。**第一处**是"只输出一个 JSON 对象"——这不是矫情。模型默认喜欢"先寒暄、再解释、最后给结果"，而解析脚本只认第一字符到最后一字符都是 JSON 的文本；一旦多出一句"好的，这是你的场景："，整段解析就失败了。**第二处**是"不要自行发明字段"——模型很乐于帮你加字段（比如自己发明一个 `mood_level`），而多出来的字段会让校验脚本报错或者被悄悄忽略，两种结果都不好。**第三处**是"无法映射时写进 notes"——留一个合法的出口，比逼着模型硬编要诚实得多。

◆ 拿到 JSON 之后：四道校验

规格表生成出来只是开始。真正决定成败的是接下来这四道关，它们全部是**确定性检查**——不看画面、不靠感觉，只跑脚本或对照清单：

1. **schema 校验。**字段名对不对、类型对不对、必填项有没有缺、枚举值是否在允许范围内。这一关解决"机器读不懂"的问题。
2. **数量与目录约束。**资产条数是否超过上限、同类资产是否过量（"树上长了 200 个灯笼"）、是否有重复 id。这一关解决"机器读得懂但读出来是灾难"的问题。
3. **玩法规则校验。**用你的相机路径当作"动线"，检查：路径上的每个机位是否能看见主体、路径有没有穿过建筑内部、起点到终点是否连通。这一关解决"能建起来但走不通"的问题。
4. **预算自洽校验。**资产条目的 max_tris 求和是否超过 budget.triangles_total；贴图数量估算是否超过 texture_mb。这一关把"性能"从口头承诺变成算术题。

⚠ **解析失败不是意外，是常态**

真实项目在文档里会直接写明失败路径：**重试若干次**，重试耗尽后**回落到规范默认值**而不是崩溃；同时给提示模板打版本号，schema 变更后让旧缓存失效。请把这三条写进你自己的作业脚本或检查清单——**凡是"生成后才可能发现"的错误，都要有兜底策略。**

∫ **深入一格 · 为什么"结构层与创作层要分开"** 选学 · 方法论

初学者常犯的一个错，是试图"一次生成一切"：让模型同时负责"把楼台放在 (x, y, z)"和"写出楼台牌匾上那句诗"。结果是两头都不讨好——坐标位置因为混在散文里而难以解析，诗句因为被塞进 JSON 字段里而写得干瘪。

更深一层的原因是：**这两类内容的"正确"标准根本不同。**坐标的正确标准是**约束满足**（在范围内、不重叠、可达），可以写成一个函数自动判真假；诗句与形制的正确标准是**文化考据与审美判断**，只能由人来判。把两套标准塞进同一次生成，就等于让一个函数同时承担"计算"和"裁决"两种职责——工程上，这几乎总是错的。

所以成熟的架构会长成图 7-3 右下角那样：左边是一个**窄而深宽而浅**的创作层（自由、开放、由人审阅）。**窄的那一层交给脚本，宽的那一层交给你自己**——这也解释了为什么本章强调"人负责审美与考据"：那不是场面话，而是分工的技术必然。

7.4 程序化生成：让规则替你"摆"场景

规格表告诉引擎"要有什么"，但"具体长什么样"还有一大块是**算出来的**，不是画出来的——这就是程序化生成（procedural generation）。它听起来高深，其实你在游戏里早就见过它：**为什么《我的世界》的地形可以无限延伸？为什么 Roguelike 游戏的每一局地牢都不一样？**因为那些地形与地牢不是美术手工摆的，而是**一段带随机数的规则**跑出来的。

程序化生成对本章有两重意义。第一，它是"地形"这一要素的主力手段——文生三维工具能给你一栋楼，给不了你一片二十万平方米的丘陵。第二，它提供了一套**"约束思维"**：好的程序化生成从来不是"随便撒"，而是**"在规则允许的范围内随机"**。这个思维会让你回到 7.3 节时更明白——为什么规格表要写得那么死板。

◆ 噪声地形：把"随机"磨成"自然的起伏"

先做一个思想实验。如果让计算机在地图上给每个点随机取一个高度，你会得到一张**碎末状**的地面：相邻两个点的高度毫无关系，像电视雪花。这不是地形，这是沙子。自然地形有一个关键特征：**相邻的地方高度相近**——山不会在你脚边突然变成深谷，它会有一个过渡。

如何在"随机"里注入"过渡"？**Perlin 噪声**（一种经典的噪声算法）给出了一套优雅的解法，用生活中的东西类比最好懂：

想象一块棋盘，每个格子的**角落**都插着一支箭头，方向随机（这叫"随机梯度向量"）。现在你要问"棋盘上某个点的地势有多高"，做法是：把该点到它所在格子四个角的**方向**算出来，与四个角上的箭头各做一次"投影"（点积，可以理解为"这两个方向有多一致"），得到四个数；再按"离得越近、权重越大"的方式把这四个数**平滑地插值**混合起来（用平滑插值函数 smoothstep，而不是直线插值，这样过渡处不会有生硬的折角）。

于是你会得到一个奇迹般的结果：**每个点的高度都是随机的，但相邻点的高度又是连续变化的**——既有随机性，又有过渡。这正是"地形"需要的质感。

还有一层认识要建立：Perlin 噪声是**伪随机**的——它不接受真正的随机源，而是由输入的坐标决定输出。这意味着**同一个输入必然得到同一个输出**。这不是缺陷，而是功能：你只要记住那个"种子"（seed）数字，就能在三天后、在另一台电脑上，**一模一样地重新长出那片地形。**

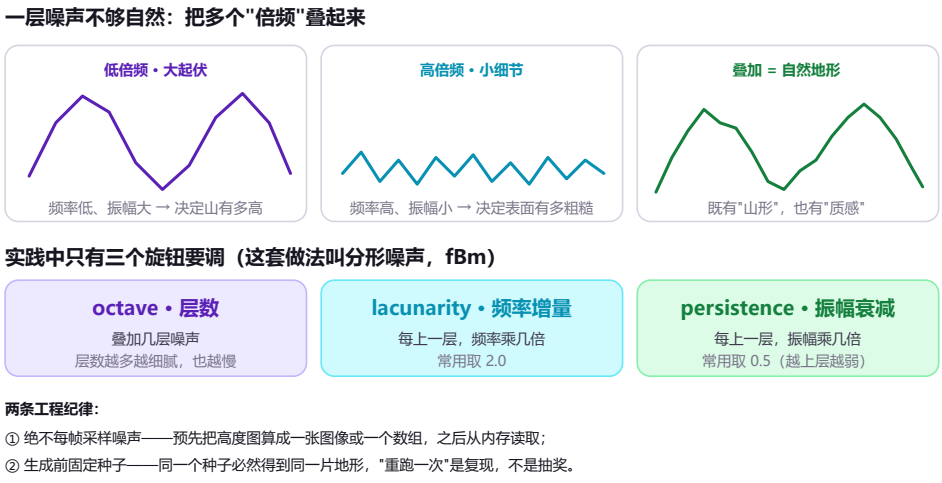


图 7-4　噪声地形的三个旋钮。程序化地形 = 低倍频定"山有几座"，高倍频定"石头有多糙"；三个参数就是全部的调音台。

在引擎里的落地方式是前文提过、这里必须点明的一句：**绝不每帧采样噪声。**噪声计算量不小，正确的做法是启动时**一次性**把高度图算好、存成一张图或一个数组（例如 513×513 的高度数组），之后每帧只是从内存里读数字。这条纪律的收益是几十倍的性能差——它是新手和专业之间最直观的一条分界线。

◆ 波函数坍缩：把"拼接"变成"推理"

第二类算法解决的是另一种需求：**我要一片看起来风格统一、但每次都不一样的瓦片地图。**比如一片古镇的院落、一座洞穴的内部、一段走廊与房间的组合。这类问题的难点不在"随机"，而在**"随便摆出来的东西拼不上"**——屋顶瓦片接不上墙，路走到一半撞墙。

波函数坍缩（Wave Function Collapse，缩写 WFC，由 Gumin 于 2016 年提出）是一个名字唬人、思想却很朴素的方法。它的做法是：

1. **准备样本。**先给一张"样例图"（或一套瓦片），把它切成许多 N×N 的小图案块（比如 3×3 的像素块）。这就是你的"素材库"与"拼接规则集"。
2. **初始化"可能性"。**新建一张空地图，每个格子都列出"所有可能出现的图案"——此时每个格子都处于"不确定态"，这叫它的**叠加态。**

3. **观察：挑最"确定"的格子先动手。**找到"剩余可选图案最少"的那个格子（也就是**熵最低**的格子），从它的候选里随机选一个，把它**坍缩**为确定值。
4. **传播：把决定通知邻居。**格子一旦确定，它的邻居就"少了一些可能"——凡是与它拼接不上的图案，都要从邻居的候选列表里删掉。删除又会继续影响邻居的邻居，如此链式传播。
5. **重复到最后一格。**全部格子坍缩完毕，整张地图就形成了：它与样例图的**局部风格一致**，但整体布局是全新的。

为什么要"先挑熵最低的"？类比一下：**考试时先做最有把握的题。**每一道题都代表一个格子的可能性集合；先做最确定的题，就能立刻把相关的题目排除掉一大片选项，越做越顺。反过来，如果先挑可能性最多的格子乱猜，一旦猜错，错误会沿着传播链扩散，最后整张地图崩坏——这就是 WFC 必须处理的"矛盾"问题：**当某个格子的候选被删空时，说明已经无解，需要回溯或重启**，同时必须设一个"最大迭代次数"的安全阀，避免程序陷入死循环。

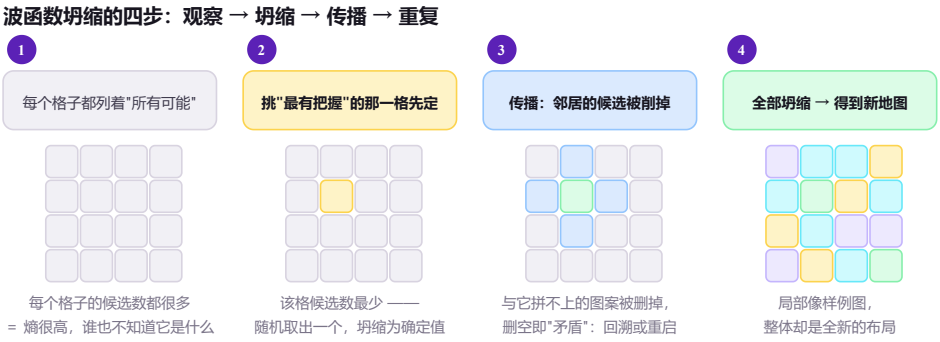


图 7-5　波函数坍缩的四步循环。绿格是已确定的格子，蓝格是"候选被削减"的邻居。整个算法的精髓是"先做最有把握的决定，然后让这个决定去影响别人"。

> 🧪 **原理注 · "熵"在这里到底指什么**
>
> 熵（entropy）在信息论里衡量"不确定性的大小"。WFC 里的"熵最低"有具体的算法含义：**看哪个格子的候选图案数量最少。**候选越少，不确定性越低，越接近"确定"。实际实现中还会加权（把出现概率也考虑进去），但教学上抓住"候选数量"这一条就够了——它解释了 WFC 的全部行为：为什么它比纯随机稳、为什么会产生矛盾、以及为什么"简单布局不该用 WFC"（**你的地图如果本来就没什么约束，用一把重锤砸一颗芝麻，反而更容易死循环**）。

◆ 泊松盘采样：让物体散布看起来"自然"

地形铺好了，接下来要往上撒东西：树、石头、草丛、灯。最直觉的做法是"每个点随机取一个坐标"——结果是**有的地方挤成一团，有的地方大片空白**；更糟的是物体会互相重叠、互相穿插，看起来像个 bug。

解决办法叫**泊松盘采样**（Poisson disk sampling）：要求任意两个点的间距不小于某个最小值，同时整体分布保持随机。它产出的分布叫"蓝噪声"——**均匀但不成网格。**生活类比：**撒种子。**你不会把种子撒成棋盘格（那不像自然），也不会随手一扬（那会长成一块块秃斑）；你会尽量让每颗种子之间保持一点点距离。标准实现是 Bridson 算法（常用参数 k=30）；想省算力，用"抖动网格"近似也行——先画规则网格，再把每个点在其格内随机挪一挪。

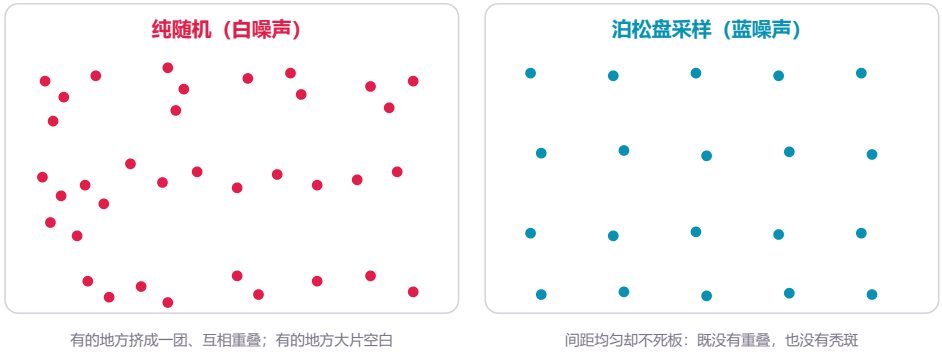


图 7-6　两种物体散布方式的对照。同一批"树"，两种撒法给出两种可信度。当你的场景"看起来像撒了一地无关的东西"时，先检查散布算法，而不是先换模型。

怎么选：一张决策表

程序化生成的算法不止一种，但它们其实各管一类需求。下面这张图是一份来自实践文档的"算法决策表"，可以直接当作选型速查：看到需求，往右找算法。

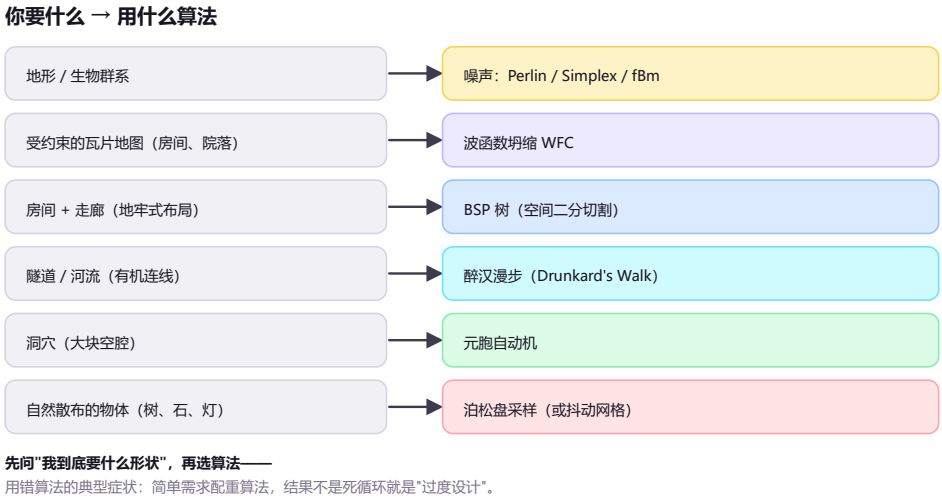


图 7-7　程序化生成的算法选型表。六种需求对应六类算法。这张表的价值在于"反向使用"：当你发现自己的生成结果怪怪的，先回来核对你选的是不是这一类需求该用的算法。

表 7-5　程序化生成算法速查（表述、用途与坑）

技术	准确表述	典型用途	注意事项
波函数坍缩（WFC，2016）	约束驱动算法：把输入拆成 N×N 图案，运行"观察—传播"循环——选剩余候选最少的格子坍缩，进而约束邻格；完全坍缩后得到与原图特性相似的新地图。另有"简单瓦片模型"变体：每块瓦片四边带标识符，限制哪些瓦片可放在相邻格	受约束的瓦片地图、风格统一的室内与洞穴布局	必须检测矛盾（无有效瓦片时回溯或重启）；设置最大迭代次数以避免死循环；**简单布局不要用复杂 WFC**（改用醉汉漫步或 BSP）
Perlin 噪声	一块格网，每个格角有一个随机梯度向量；对每个点，计算"到格角的距离向量"与"格角梯度"的点积，再用平滑插值混合。噪声是**伪随机**的——相同输入得到相同输出	二维高度图、三维点云、体素世界的属性分布	**性能关键：绝不每帧采样噪声**——预生成一次高度图存进图像或数组，此后从内存读取
分形噪声（fBm）	跨倍频累积任意噪声类型，三个参数控制：层数（octave）、频率增量（lacunarity）、振幅衰减（persistence）	更自然的地形层次	参数直接写进规格表，才能复现同一片地形
泊松盘采样	提供"均匀但非网格"的物体放置。纯随机（白噪声）会造成聚集与重叠；本方法或抖动网格给出自然分布	放置物体、敌人、植被	标准做法是 Bridson 算法（k=30，蓝噪声分布）

∫ 深入一格 · 伪随机为什么比"真随机"更适合创作 选学 · 方法论

"伪随机"听起来像缺点：既然是假的，为什么不用真随机？因为在创作流程里，**可复现比不可预测重要得多。**

先看真随机会带来什么。如果地形由真随机决定，那么你今天生成了一个满意的江岸，明天想再生成一次"同一个江岸但树多一点"，是做不到的——你只能重新抽一次奖，而抽出来的东西和昨天毫无关系。整个"迭代改进"的流程就断了：你没有"版本"，只有"另一次开奖"。

伪随机解决的正是这件事：它以种子为输入，输出一段看起来完全随机的序列；**同一个种子给出同一段序列。**于是"种子"就成了你的作品版本号：`seed=20261009` 是"那版江岸"，换一个种子就是"再开一局"。这也是为什么程序化生成的工程原则里，第一条写的是"**始终使用带种子的伪随机数生成器以保证可复现**"。

把这条原则翻译成创作语言，就是本章反复出现的那句：**把一切能变成数字的东西变成数字。**种子里放数字，参数写成数字，资产编号写成数字——只有可复现的东西，才谈得上打磨。

🎯 三个跨技术通用原则

1. **始终使用带种子的伪随机数生成器**，保证可复现（常见的做法是使用短周期的高质量随机数算法，如 mulberry32、sfc32 这类）。
2. **校验生成的输出**——例如从起点到终点跑一次寻路，确认真的走通；程序化生成的产物在"看起来还行"和"真的可用"之间，差距往往就在这里。
3. **大世界把生成移出主线程**——分块、多线程生成，避免玩家在"卡顿的空白世界"里等地图出现。

7.5 资产生产管线：每一类资产有它自己的"生产线"

规格表定好了要什么，程序化算法铺好了地形，接下来是真正的"造东西"环节。这一节的核心观念是：**不要指望一个工具包打天下。**天空盒、地形、植被、建筑、道具、材质，各自有各自的生成路径、各自的输出格式、各自的验收硬指标。把它们排成一条有序的流水线，比同时开十个网页靠谱得多。

图 7-8　场景资产生产管线的五个阶段。顺序不能乱：环境决定光，光决定物体的观感；反过来做，就得反复返工。

天空盒：一个工具解决的"整个天空"

天空盒（skybox）是包裹住整个场景的 360° 环境图，它同时干三件事：提供天空的视觉内容、为场景提供环境光（间接照明的来源）、以及给水面和金属表面提供反射的"内容"。可以说，天空盒是"氛围"这一要素的技术载体。

它的生成方式已经相当成熟，检索到的最具代表性的工具是 **Skybox AI**（Blockade Labs）。它把一段文本提示转成生产级的 **360° 等距柱状**（equirectangular）环境图，秒级完成；官方自述已服务 150 万以上创作者、生成超过 1000 万个天空盒。它的工作流很简单，值得完整记下来：

1. **写提示词描述环境。**对场景、光照、时刻、氛围描述得越具体越好——例如"满月之下的开阔江面，薄云，远处山影，冷蓝主调"。
2. **选风格包预设。**平台提供 30 余种精选风格包，覆盖从照片级到动漫、数字绘画、低多边形、科幻等取向。**这一步是"视觉统一"最省力的抓手**：全组用同一个风格包，天空的"画法"就统一了。
3. **生成并环视。**数秒出图，可以点击拖拽做 360° 环视检查；不满意就重新生成或微调提示词。
4. **下载。**导出选项包括 1K 到 8K（默认 8K，即 8192×4096，最高支持 16K）等分辨率，格式上有等距柱状图、立方体贴图、HDR/EXR、深度图，甚至三维网格格式。
5. **导入引擎。**通常拖放即可设为场景的环境或背景贴图。

它的能力边界也要写清楚，免得你在错误的场景里浪费时间：官方明确它**不设计用于创建人或动物**；当前**不能创建"引用其他天空盒"的天空盒**来实现连贯的连通空间（该能力在开发中）；也不支持等距柱状视频导出。免费档有明显限制（例如每月约 15 个天空盒，且带水印、按 CC-BY 许可使用），付费档才解锁更高配额、优先速度、商业许可与高级导出。**价格、额度与版本在 2026 年变动频繁，使用前请以官方页面为准**——这是本书对所有工具类信息的一贯要求。

✎ 提示词 · 夜江天空盒（360°）

夜色下的开阔江面与远山，**[一轮满月]**位于画面上方，**[薄云]**透出微光；地平线处有淡淡水汽，远处山影层叠；整体**[冷蓝]**色调，月光为主要光源；**[中国水墨动画]**风格，低对比、无人物、无动物、无文字；等距柱状全景，左右边界需无缝衔接

风格包 水墨 / 数字绘画　分辨率 8K　导出 HDR + 深度图　负向 人物/动物/文字

这段提示词里有两个容易被忽略的细节，值得单独指出。**第一是"左右边界需无缝衔接"**：等距柱状图是把球面展开成矩形，如果你不强调，生成结果左右两条边往往接不上，环绕一圈时会看到一道明显的"接缝"。**第二是"无人物、无动物、无文字"**：天空是背景，一旦生成出一只飞鸟或一行文字，它会在环绕时被拉伸变形，成为画面上的笑话。把"不要什么"写进负向提示，是专业提示词与业余提示词最常见的区别之一。

◆ 地形、水面与植被：三种"量很大但很便宜"的资产

这三类资产有共同特点：**占地大、数量多、但单件面数很低。**它们是性能预算里最容易被浪费的地方，也是最容易通过工程手段省下来的地方。

地形由 7.4 节的噪声算法生成高度，再由你手工"雕刻"关键区域——把楼台所在的那块地压平、把江岸修出一条缓坡。这一步的价值在于：**算法给出的是"自然"，人给出的是"可用"**。程序化地形能长出漂亮的山，但不会主动给你一块能放下房子的平地。

水面看似简单，其实是最容易露馅的一层。技术上它通常就是"一个平面 + 一套专用材质"——面数极低（两三个三角面就够），但材质里包含了反射、折射、波纹法线等一堆计算。它的验收要点有两句话：**反射的方向必须与光照一致**（月亮在高处，水里的月亮倒影就要相应地长而碎，不能是一轮完整的圆），**水面高度必须与地形统一坐标系**（否则会出现"船浮在山坡上"这种经典 bug）。

植被的正道是"低模 + 实例化 + 程序化散布"：不要造 200 棵各不相同的高模树，而要造 2–3 种低模树，然后用 7.4 节的泊松盘采样把它们散布出去，并在引擎里以**实例**的方式重复使用。这样 200 棵树只占一份模型内存——这正是降低绘制调用的核心技巧。

◆ 建筑与道具：文生三维真正的主场

建筑与道具是"文生三维"能力最直接的体现：输入一段描述（或一张参考图），输出一个三维网格。这里有三条来自实践的经验：**其一，小物件成功率最高**——扁舟、灯笼、酒樽这类体量小、结构清晰的物体，生成质量明显优于复杂的建筑群。所以教学与原型阶段，建议从道具练手。**其二，建筑要"先定体量、后定细节"**——先确认它的长宽高与占地是否与地形、与镜头匹配，再去抠飞檐斗拱；顺序反了，抠得再精细也会被整体比例推翻。**其三，参考图比文字更可靠**——"唐风楼台"这四个字在不同模型脑中的画面差异巨大，而一张明确的参考图能把这个差异压缩掉大半。

这也顺带解释了第 4 章讲过的那条结论为什么重要：生成出来的网格**不一定拓扑干净**——可能有非流形边、内部面、反转法线，UV 也常常混乱（重叠、缺失、纹素密度不一致）。所以本章把"导出前检查"（7.6 节）视为管线里的必经环节，而不是可选项。第 4 章已详细讲过拓扑与 UV 的处理，这里只强调它在场景里的特殊后果：**一个道具的 UV 乱了，你损失的是一张贴图；一片建筑的 UV 乱了，你损失的是整个画面的统一感。**

◆ 材质：一次把话说清楚的"表面说明书"

材质（material）决定"这个表面怎么与光互动"。现代引擎使用 **PBR**（物理基础渲染）工作流，把表面属性拆到多张贴图上、共享同一套 UV 布局。对场景创作来说，只要记住三类贴图的分工即可：

- **基础色贴图（Albedo / Base Color）**——纯表面颜色，**不含烘焙的阴影与高光。**这一条是初学者最容易违反的：如果你把一张"带着好看阴影"的贴图直接贴上去，一旦场景光照变化，那种阴影就会与真实光影打架，看起来"脏"。
- **法线贴图（Normal map）**——用 RGB 三个通道编码表面凹凸的方向偏转，**不增加多边形就能让平面看起来有细节。**砖缝、木纹、瓦楞的"手感"基本都靠它。
- **粗糙度与金属度（Roughness / Metallic）**——控制反射的锐利程度与"这是金属还是非金属"。粗糙度 0 相当于镜面，1 相当于完全哑光。

材质环节最大的坑不在生成，而在**引擎之间的约定差异。**这里有三条必须记住的"翻译规则"：Unity 暴露的是 **平滑度 = 1 − 粗糙度**（外部导出的粗糙度贴图要用它的反相）；Unity 默认是 Y 轴向上，而 Blender 是 Z 轴向上；Unity 与 Unreal 之间**法线贴图的绿通道朝向相反。**还有一条更隐蔽的：数字型（非颜色）贴图必须**关闭 sRGB 色彩空间**，否则你的粗糙度会被引擎按"颜色"再解释一遍，反射全乱。这些差异就是"看起来对"和"真的对"之间的那条分界线。

表 7-6　六类资产的生成路径、输出格式与验收硬指标

资产类型	主要生成路径	输出格式	提示词要点	验收硬指标
天空盒	文生 360° 等距柱状环境图	HDR / EXR / 立方体贴图 / 等距柱状 PNG	时刻 + 光照方向 + 云量 + 色调；强调无缝、无人物动物文字	左右边界能接上；分辨率与色深满足后续光照需求
地形	程序化噪声生成 + 手工雕刻关键区域	高度图 / 引擎地形数据	把噪声参数与种子写进规格表	坡度可通行；与水面共用坐标系；建筑落位处是平地
水面	平面网格 + 专用水体材质	材质 + 极少量网格	反射来源指向天空盒；波纹尺度与镜头距离匹配	反射方向与主光源一致；不穿透地形、不悬浮
植被	低模资产 + 泊松盘散布 + 实例化	网格 + 引擎植被/实例系统	统一树型与色彩；数量受目录约束限制	实例数量不超上限；无重叠穿插；远景有 LOD 替代
建筑	文生三维 / 图生三维（带参考图）	GLB / FBX	先写体量与占地，再写形制与时代特征；附参考图	面数达标、Pivot 合理、UV 可用、比例与参照物一致
道具	文生三维（小物件成功率最高）	GLB / FBX	说明用途、尺寸与材质直觉（木质/金属/织物）	面数达标；Pivot 在几何中心（便于旋转摆放）

> 📌 **跨章连接：这里用到的第 4、5、6 章成果**
>
> 本章的资产不是凭空来的——楼台与扁舟来自第 4 章的三维资产生成与导出质检，诗人角色来自第 5 章的角色动画，舞者/捣衣动作来自第 6 章的姿态迁移。本章的贡献是**给出"它们放在哪里、多大、朝向如何"的那张地图**，以及"拼起来之后还能跑"的那几道检查。

7.6 落进引擎：导入流程、常见坑与性能预算

现在到了最容易翻车、也最有教学价值的一节。资产在生成工具里看起来完美，拖进引擎却"横着躺、大一百倍、颜色发粉"——这不是玄学，而是四个具体的技术原因：**单位、变换、轴向、Pivot**。把它们一次性讲清楚，你以后能省下大量瞎调的时间。

◆ 四个"不可见"的坑

第一个坑是尺度（unit / scale）。检索到的工程资料把它称为"最常见的失败点"：AI 模型常以"任意真实世界尺度"导出，于是出现经典的"我的模型在哪？"问题。最典型的意外是**单位不匹配**——按米创作的模型，如果导入器假设是厘米，就会大或小 100 倍。**Unity 用米，Unreal 用厘米**，这句话请你记牢。处理办法：导出前用已知单位（例如约定 1 单位 = 1 米）对照原始人形或立方体参考缩放；在 Unity 里于资产的导入设置中设置缩放因子以匹配场景单位；在 Blender 里**应用变换**，让 scale 读数变成 (1, 1, 1) 再导出。

第二个坑是"未应用变换"。如果一个资产身上挂着一个 scale 为 (1.0, 1.0, 0.01) 且没有应用过的变换，它"**到达引擎时会变成一个煎饼**"。原因回到 7.1 节的层级数学：这个缩放会与父级、与引擎的导入设置相乘，结果无法预料。所以规矩是死的：**导出前应用全部变换。**

第三个坑是轴向（axis）。Blender 是 Z 轴向上；Unity 和 Unreal 是 Y 轴向上。FBX 与 glTF 这类格式会记录轴信息，但**不匹配仍会导致模型翻转或旋转**——预览中完美的模型可能"横着到达"（这正是小舟遇到的那一幕）。处理办法：在建模工具里把朝向修正为直立并应用旋转，导入引擎后再验证一次；资产应当对齐引擎的前向轴。

第四个坑是 Pivot / 原点。偏离中心的 Pivot"**让引擎内放置成为噩梦**"：它可能落在世界原点、远在网格之外，或者出现"原点漂移"与包围盒不匹配——**这些在渲染截图里完全看不见，却会在场景里破坏放置**。正确的做法是**按逻辑设置 Pivot**：环境资产在地平面上，角色在颈根，机械物体在几何中心。

> ∫ **深入一格 · 为什么"截图看起来对"证明不了任何事** 选学 · 方法论
>
> 这是本章最反直觉、也最值得记住的一条工程观点：**截图或精美的转台动画只证明了外观，它不证明尺度、不证明变换是否应用、不证明 Pivot 是否合理、不证明有没有非流形边，也不证明多边形预算是否达标。**
>
> 为什么？因为上述四样东西**都是"在某个特定上下文里才显形"的属性。**一张孤立渲染的模型图，没有参照物，所以你看不出它是不是 100 倍大；没有坐标系，所以你看不出它是站是躺；没有操作，所以你看不出 Pivot 在哪；没有统计面板，所以你看不出它有多少面。它们就像**一张身份证照片上的信息**——照片能告诉你这个人长什么样，但告诉不了你他的身高、国籍和血型。
>
> 所以专业的做法是改用**确定性断言**（deterministic assertion）：不看图，而是**读数字。**例如检查资产是否已应用缩放（一个断言问"scale 是否读作 1,1,1"）、检查模型是否存在非流形边（断言问"非流形边计数是否为 0"），以及最朴素也最有效的一招——**导出回读**：导出之后重新读回来，验证数据与导出前一致。这套思路可以概括成一句工程结论，值得抄在你的工作笔记第一页：**"引擎导入决定资产能否出货。"**验证必须在导出前完成——引擎导入是最终检查，不是调试阶段。

图 7-9　导出前检查表。左四右四共八项，覆盖几何、UV、材质、格式、尺度与 Pivot。工程资料的经验估计是：一份这样的检查表能捕获约九成会破坏引擎导入的问题。

表 7-7　AI 资产进引擎的常见坑：表现、后果与处理

坑	具体表现（工程资料中的原话）	处理
尺度	"最常见的失败点"。AI 模型常以"任意真实世界尺度"导出，导致经典的"我的模型在哪？"问题；单位不匹配是最常见的意外——按米创作的模型若导入器假设厘米，会大或小 100 倍。Unity 用米，Unreal 用厘米	导出前用已知单位对照原始人形或立方体参考缩放；在引擎资产导入设置里设好缩放因子；在建模工具中应用变换，使缩放读作 (1, 1, 1)
未应用变换	带缩放 (1.0, 1.0, 0.01) 且未应用变换的资产"到达 Unreal 会变成一个煎饼"	导出前应用全部变换
轴向	建模工具常见 Z 轴向上，Unity 与 Unreal 是 Y 轴向上。格式会记录轴信息，但不匹配仍会导致模型翻转或旋转——预览中完美的模型可能"横着到达"	在建模工具里修正朝向（旋转到直立并应用旋转），再在引擎里验证；资产应对齐引擎的前向轴
Pivot / 原点	偏离中心的 Pivot"让引擎内放置成为噩梦"；它可能在世界原点、远离网格之外，或出现"原点漂移"与包围盒不匹配——渲染中不可见，却破坏场景放置	按逻辑设置 Pivot：环境资产在地平面上、角色在颈根、机械物体在几何中心
面数与拓扑	过密三角网格、混乱三角、不一致边流、非流形边、内部面、游离顶点、空洞、反转法线，会导致着色错误或导入失败	**会形变的物体（角色、生物、关节道具）必须重拓扑**（以干净的四边形为主、边流可预测）；**静态非形变道具可接受减面处理**
UV	AI 生成的 UV 常混乱——重叠壳、缺失 UV、纹素密度不一致——导致纹理损坏	重新生成或修复 UV，让岛屿适配 0–1 空间并留出间距。**注意：重拓扑必须先于 UV 展开**
材质与贴图	缺失贴图、断裂引用、PBR 数据"在引擎间不总能干净映射"；Unity 里的粉色材质通常表示材质与着色器的交接问题，不是重新导入能解决的；法线贴图绿通道朝向在 Unity 与 Unreal 间不同	逐张贴图验证通道与色彩空间；数据贴图关闭 sRGB
格式选择	FBX 用于复杂高保真的引擎项目（携带网格、UV、材质、骨架、动画）；glTF/GLB 适合 Web 与移动，是"友好的默认"——把几何、PBR 材质与贴图打进一个文件，但其材质系统在引擎中可能需要转换	按目标平台选择：把 GLB 当"预览与交付"格式，把 FBX 当"引擎工作"格式
碰撞与 LOD	标准的 AI 导出**不含经过验证的碰撞网格**	生成简化碰撞体（盒、胶囊、凸包）；为重复或远处的资产生成 LOD 链

◆ 性能预算：把"会不会卡"变成一道算术题

性能不是一个"跑起来才知道"的玄学问题，而是一份**预算表**。它的结构很像家庭记账：先定总收入（目标平台能承受的开销），再给每一类资产分配额度，最后逐项核对支出。

先看最直观的一笔账——**三角面预算**。下面这张图是一次工程资料的汇总，注意刻度是对数的：从几百面到几十万面跨了三个数量级，线性刻度根本画不下。

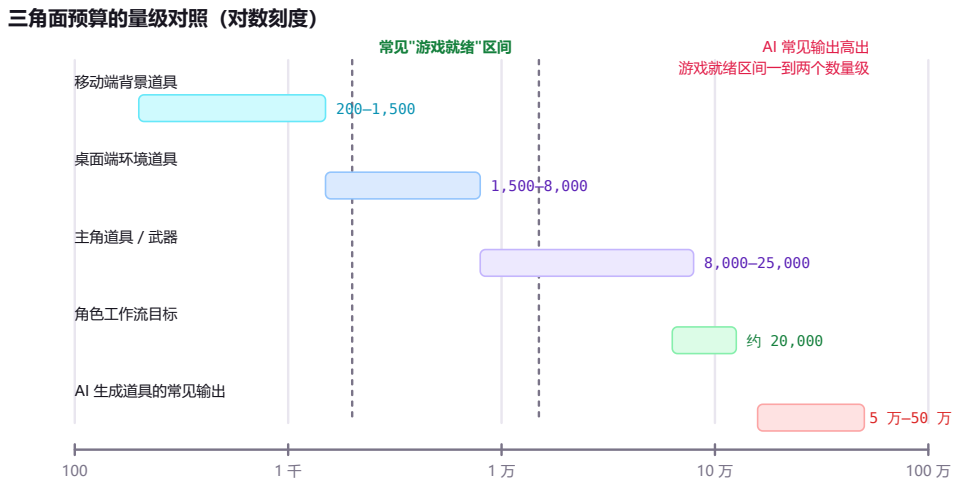


图 7-10　面数预算的量级关系。横轴是对数刻度。绿色虚线之间是常见的"游戏就绪"区间；最下方一根红条是 AI 生成道具的常见输出区间——它高出就绪区间一到两个数量级，这正是"面数爆炸"问题的量级来源。

表 7-8　三角面预算参考（按资产类型划分）

资产类型	三角面预算	说明
移动端背景道具	200–1,500	画面最远、最不起眼的一层，预算最紧
桌面端环境道具	1,500–8,000	近景可辨识的物件，如桌案、灯笼、石灯
主角道具 / 武器	8,000–25,000	会被特写、被拿在手里的物件，预算最高
角色工作流目标	约 20,000	角色是"要被看清"的主体，且需形变边环
某三维生成工具的默认网格档	约 5,000	作为对照：工具默认值不等于项目预算

第二笔账更隐蔽，但往往更早成为瓶颈：**贴图内存。**工程资料明确指出，UE5 的 Nanite 等虚拟几何技术虽然放松了传统多边形数量的约束，**但拓扑、UV 质量、贴图内存与材质复杂度仍然影响性能——而且贴图内存通常比多边形数更早成为瓶颈。**这句话值得反复读：很多人把注意力全放在"减面"上，真正的瓶颈却在"你有多少张 4K 贴图"。

第三笔账是 **LOD**（细节层级，level of detail）。原理很朴素：**远处的物体不需要那么多面。**同一棵树，近处用高模、中距离用中模、远距离用一张"贴着树形状的纸片"，投影到屏幕上几乎没有差别，开销却差了几十倍。工程资料的建议是明确的：**重复出现的资产与远处的资产都需要 LOD 链。**

第四笔账是**绘制调用（draw call）**。它是"引擎向显卡提交一次绘制命令"的动作。这里要特别说明本书的诚实：**检索范围内没有拿到 2026 年的绘制调用预算标准数字**，所以这里只讲原理，不给配额——**合并材质、减少材质数量、对重复物体使用实例化，是降低绘制调用的三个基本手段。**注意它与 LOD 的关系：把 200 棵独立的树变成"1 个模型 + 200 个实例"，省下来的主要不是内存，而是调用次数。

该省的省

远景资产的模型精度、看不到的背面、重复物体的独立材质、超高分辨率的贴图、法线强度过高的细节。这些"省下来"几乎无损观感。

不该省的地方

主光源方向的统一、水面反射与光源的一致性、主体资产的 UV 与纹素密度、角色形变区的边环——这些一省就"露馅"。

先做减法，再做加法

先把资产数量与面数压到预算内跑通一遍"能看的场景"，再往上加细节。反过来（先堆满再加）几乎必然要推翻重做。

把预算写进规格表

预算是约束，不是事后统计。把它写进 7.3 节的 JSON（`max_tris`、`budget.triangles_total`），生成阶段就有红线可依。

7.7 关键概念表

本章术语密度较高，这里集中落表，作为复习与查检之用。建议对照你自己的工作文件逐个打勾——每一个词，都对应一次你会在实践里真正遇到的判断。

表 7-9　第 7 章关键概念（中英对照）

中文	英文	一句话解释
场景	scene	由地形、建筑、道具、光照、氛围构成的、可被观看或行走的三维空间
场景规格表	scene spec / scene specification	把自然语言需求转写成字段化数据的清单，通常是 JSON
结构化输出	structured output	要求模型按规定格式（如 JSON schema）作答，而不是自由文本
schema	schema	字段名、类型与取值范围的"表格规范"，用于自动校验
校验	validation	用脚本或规则检查数据是否合法；分 schema 校验与玩法规则校验两层
目录约束	catalog limits	对资产数量、类型与规模的硬上限，防止"生成 500 棵树"
连通性	connectivity	起点到终点是否真的存在通路，是玩法校验的第一条
可解性	solvability	关卡是否存在解法；成熟工作流坚持"先设计解法路径，再写地图"
程序化生成	procedural generation	用带随机数的规则算出内容，而不是手工摆放每一件
噪声	noise	可产生连续随机起伏的算法；Perlin 噪声是其经典代表
分形噪声	fractal noise, fBm	把多层不同频率的噪声叠加起来，得到更自然的地形
倍频	octave	fBm 叠加的层数；层数越多细节越丰富
频率增量 / 振幅衰减	lacunarity / persistence	每上一层的频率倍数与振幅倍数，常用 2.0 与 0.5
波函数坍缩	wave function collapse, WFC	用"观察—传播"循环生成与样例风格一致的新布局
熵	entropy	不确定性的度量；WFC 里指候选图案的数量
泊松盘采样	Poisson disk sampling	给出"均匀但非网格"的物体分布，避免聚集与重叠
蓝噪声	blue noise	泊松盘采样产出的分布类型，视觉上最接近自然散布
种子	seed	伪随机的输入；同一种子必然得到同一结果，是复现的基础
天空盒	skybox	包裹整个场景的 360° 环境图，同时提供天空内容与环境光
等距柱状投影	equirectangular projection	把球面环境展成矩形的投影方式，天空盒的通用格式
地形	terrain	场景的地面起伏与地表材质，通常由高度场表示
资产管线	asset pipeline	从需求到可导入资产的加工流程，含生成、检查、导出、导入
网格	mesh	三维物体的几何表示，由顶点、边与面构成
拓扑	topology	网格的连接结构；决定能否形变、能否干净展开 UV
重拓扑	retopology	把高密度混乱三角网格重做成以四边形为主、边流可预测的网格
非流形边	non-manifold edge	被三个以上面共享的边；会导致着色错误或导入失败
水密	watertight	模型表面完全闭合、没有孔洞；打印与仿真必需，渲染则视需求而定
UV 展开	UV unwrapping	把三维表面投影到二维坐标，使贴图能正确落位
纹素密度	texel density	贴图在模型表面的像素密度；不一致会显得有的地方糊、有的地方太锐
法线贴图	normal map	用 RGB 编码表面凹凸方向，不增加多边形就产生细节
粗糙度 / 金属度	roughness / metallic	控制反射锐度与"金属还是非金属"的 PBR 参数
物理基础渲染	PBR	把表面属性拆成一整套贴图、遵循能量守恒的渲染工作流
面数	poly count	网格的三角面数量，性能预算里最直观的一笔账

中文	英文	一句话解释
细节层级	level of detail, LOD	按距离切换不同精度的模型，远处用低模
绘制调用	draw call	引擎向显卡提交一次绘制命令；合并材质与实例化可降低它
实例化	instancing	同一个模型被多次复用但只占一份资源，专治"两百棵树"
碰撞体	collision mesh	供物理引擎使用的简化几何；AI 导出通常不包含它
变换 / 应用变换	transform / apply transform	物体的位置、旋转与缩放；把缩放"固化"进网格再导出，是导入成功的前提
轴心点	pivot	物体的原点与旋转中心；它的位置决定场景内放置的手感
轴向	axis	哪个轴朝上；建模工具与引擎不一致会导致模型"横着到达"
层级	hierarchy	场景里物体的父子树结构；父级移动会带着子级一起动
动线	flow / pathing	观众或玩家在场景中移动的路径，决定场景"为谁而建"
确定性断言	deterministic assertion	不看图、只读数字的检查，如"缩放是否已应用""是否存在非流形边"
导出回读	export round-trip	导出后重新读回文件，验证数据与导出前一致
级联误差	cascading error	多步骤生成中每一步的小偏差累积成整体跑偏

7.8 案例教学：三支团队，一条共同的流水线

接下来三个案例，规模从"三个人一学期"到"上千人的工业化管线"，但你会发现它们踩的是同一块地板：**先把需求结构化，再让工具批量执行，最后靠人去检查和判断"对不对"**。这个共性，就是我们这个通识课要教给你的东西。

> 🩺 **案例 7-A** § 春江小组的"夜江场景"，从一句话到十二件资产
>
> **背景：**第 9 周，三人小组要为《春江花月夜》短片搭建主场景。小江的需求原话只有一句："夜里的江边，一轮明月，右岸有一座临江楼台，水面上漂着一条小船，月色清冷。"上一周，小舟已经在引擎里被这句"文学描述"折腾得够呛。
>
> **做法：**这次他们决定先做清单。小月先定参考图与色板（5 张水墨夜景 + 5 个十六进制色值，写在小组共享文档第一行）；小舟用表 7-4 的字段模板写了一份提示词，让大语言模型把这句话翻译成 JSON；翻译出来是 12 件资产，他逐条核对后做了三处修改——把建筑的面数上限从模型默认的"不填"改成 8,000，把"灯笼"的位置从"楼台内部"改到画面右三分线附近（因为镜头在那儿），并把水面高度显式写成 0，与地形共用坐标系。随后是四道校验：schema 合规、数量未超上限、相机路径能从岸边平移到江心且中途不穿墙、预算求和低于 12 万面。全部通过，才开始生成。
>
> **结果：**天空盒与地形半天完成；楼台第一次生成出来是个"二层小洋楼加盖飞檐"的混搭体，换了参考图重做一次才合格；扁舟一次通过；植被用泊松盘散布 60 余株低模柳树并做了实例化。全部导入后首次运行掉帧严重，查出两个原因：楼台贴图是 4K、且重复资产没有被实例化。压到 2K 并开启实例化后，场景在教室的笔记本上稳定可演示。**整个场景初稿耗时约两天，其中"生成"占了不到三分之一，剩下时间几乎都花在校验与检查上。**
>
> **复盘：**小舟在小组日志里写了一句总结，值得抄下来：**"省下来的不是做资产的时间，而是反复返工的时间。"**他们真正的提速点在两处：一是需求被写成了可核对的数字，返工从"全部重做"变成"改一行的值"；二是环境层先于物体层完成，所以物体的受光和色调一次就对上了，没有出现"做完楼台才发现整体太亮"的连锁返工。

🩺 **案例 7-B** § 盛趣游戏：一个场景复刻只用"1 + 1 + 0.5"

背景：据公开报道，盛趣游戏在完成某款头部国产游戏（以中国神话与古建场景闻名）的局部场景复刻时，投入的人力是**1 名三维艺术家、1 位场景设计师，以及半个技术支持**；此后同类流程被应用到另一款项目的三维模型制作中，效率较纯人工建模提升约 100%。

做法：把这件事拆开看，就能看出本章讲的流水线在工业环境里的样子：场景设计师负责"需求结构化"（哪些资产、什么风格、什么预算），三维艺术家负责"生成物的甄别与重做"（哪些能直接用、哪些要重拓扑或重做 UV），技术支持则把"校验"这一步脚本化——面数统计、命名规范检查、导入报错监控。三个人各管一段，恰好对应 7.3 节那张图的三个环节。

结果：据该企业公开披露的口径，效率较纯人工提升约 100%。需要清楚说明的是：**这类效率数字来自企业公开信息与媒体报道，口径各异，不是第三方测量的客观结论**，教材引用它，是为说明"分工结构"而非为说明"AI 有多强"。

复盘：这个案例最值得学的是**人数结构**：三个人里没有一个专职"操作生成工具"的岗位。这说明在成熟的管线里，"生成"已经变成了一个按钮，真正的工作量转移到了**需求定义、资产甄别与流程自动化**三处——这也正是本章希望你建立的能力结构。原场景是实拍级的实景复原，也提示了一条实用路径：**有真实参照物时，优先用参照（照片、测量、扫描）而不是纯想象**，它同时解决"像不像"和"对不对"两个问题。

🩺 **案例 7-C** § 《逆水寒》：效率提升 70% 的背后，是一套"风格资产库"

背景：据公开报道与媒体报道，网易在《逆水寒》项目中通过 AI 原生管线覆盖原画、模型、动画、音频、关卡、测试等环节，其中**场景设计效率提升约 70%**，动画产能综合提效 30%–300%，AI 生成任务脚本的时间从 2 周缩短到 3 天、人力成本下降约 40%。同批公开信息里还有一个细节更值得注意：其美术资源平台可以对具体项目做**一键式风格化模型微调**，目标是"确保生成内容与项目美术设定统一"。

做法：关键就在那句"确保风格统一"。大项目不可能靠"每次写好提示词"来维持风格——那太不稳定。它们的做法是把风格**沉淀成资产**：把项目的美术设定训练/微调成一个专用模型，让所有生成都从同一个"风格原点"出发。这与 7.2 节给学生的土办法（参考图集 + 色板 + 材质命名规范）是同一个道理，只是量级不同：**学生用一套图统一风格，工业管线用一套模型统一风格。**

结果：据企业公开披露，部分环节效能提升显著。同样需要注明：这些数字属于企业自称或媒体转述，口径各异；公开信息里同时提到该管线运行在支持 1 万–3 万节点的高并发私有云上——**"效率提升"的背后是巨大的基础设施投入**，这一点常被忽略。

复盘：这个案例给通识课读者的启示不是"大厂效率高"，而是**"风格"是可以被工程化的对象。**你不需要私有云，但你完全可以做三件同构的事：建一个属于本组的参考图集、固定一套色板、给材质起统一的命名规则。这三件事的成本几乎为零，收益却是"全片看起来像一个人做的"。

表 7-10　本章涉及的行业效率数据（引用时必须注明来源性质）

数据	来源性质	可用来说明什么
国内收入排名前 50 的游戏企业已有约 80% 公开披露开展 AI 相关布局；游戏企业 AI 技术普及率约 86%	行业报告与协会负责人公开发言（伽马数据 2026 年 7 月底报告、中国音数协）	AI 进入场景与美术生产已是行业常态，不是实验
AI 美术提效 30%–300%；美术设计是 AI 应用最广环节（约 84.2%）；三维资产生成从"数周压缩至数分钟"	券商研究报告测算	区间极宽，说明"效率"高度依赖流程与资产类型
某企业全链路 AI 创作方案应用于近 100 个游戏项目，部分生产环节效率最高提升约 8 倍；角色动画制作从 3 天压缩到 15 分钟	企业公开披露（媒体报道转述）	"部分环节"与"最高"是关键限定词，不可当作整条管线均值
三维资产生产从"5–7 天、数千元"变为"10 分钟、数元"	企业云业务公开宣传	单件资产的生成成本确实坍缩，但不计人工检查与返工
某城市布局工具把 25 平方千米虚拟城市的开发周期从 5 天缩短至 25 分钟	企业公开披露	程序化 + 结构化的组合，在大规模重复内容上收益最大
某项目局部场景复刻投入 1 名三维艺术家 + 1 位场景设计师 + 半个技术支持，效率较纯人工提升约 100%	企业公开信息与媒体报道	人力结构的重组，比"用不用 AI"更能解释效率来源

⚠️ **读效率数字的三条纪律**

第一，先看口径："效率提升"说的是单件资产生成时间、整个环节的工时，还是项目周期？三者差别巨大。**第二，先看分母：**从 3 天压到 15 分钟的前提往往是"这件工作已经被高度标准化"，无法标准化的工作压不下来。**第三，看有没有算入返工：**这正是下一节失败的起点——**如果省下的时间被返工吃掉，账面上的效率就没有意义。**

7.9 失败案例诊所：三个真实的翻车现场

三个诊断案例，分别对应**性能、视觉、文化**三类事故。每个都按"症状 → 成因 → 处方"写，处方都可以直接用。

🩺 **失败诊所 F7-1** 面数爆炸：两份"看起来差不多"的场景，差了 111 倍

症状：场景在引擎里根本跑不动。小月把两份场景的截图并排给同学看，大家一致认为"差别不大"；但打开统计面板，一份是 29 万三角面，另一份是 **3278 万**三角面——差了约 111 倍。

成因：这不是某个参数的锅，而是**"生成物的默认值"与"生产物的预算"根本不是同一套标准**。检索到的一项对照案例（一项以"风格化奇幻酒馆场景"为对象、n=1 的实践导向比较案例研究）给出了具体数字：手工搭建的场景 **123 个物体、293,828 三角面**；AI 辅助搭建的场景 **103 个物体、32,781,505 三角面**，约 **111.6 倍**；建模耗时分别是 **716.06 分钟**与 **238 分钟**，差值约 **478 分钟**。报告的原话是：省下的时间"伴随着可观的**技术债**"——那 478 分钟没有消失，它转化成了后面的返工与修复。为什么生成器会输出这么密的网格？一份来自三维生成工具厂商的技术文章说得很直白：这类网格"通常带有数十万个近似均匀的三角面——对渲染、编辑和动画管线而言过于密集"；而简化算法虽然能减面，但作为纯粹的几何后处理，它产生的三角化"不尊重尖锐特征与部件结构"，仍然远达不到生产资产所需的"干净的、美术师风格的拓扑"。

处方：四步。**第一步，把面数预算写进规格表**（表 7-8 给了分档参考：移动端背景道具 200–1,500、桌面端环境道具 1,500–8,000、主角道具 8,000–25,000），生成前就设上限，而不是生成后统计。**第二步，按用途分流**——会形变的（角色、生物、关节道具）必须重拓扑，静态道具可以接受减面；工程资料给出的工时参考是中等道具约 30–60 分钟、角色 2–4 小时，请把这笔时间**计入项目排期**。**第三步，堵住贴图这条更早的瓶颈**——记住那句工程结论：贴图内存通常比多边形数更早成为瓶颈。**第四步，改验收方式**——不看转台图，读统计数字。

一句提醒：这项对照案例原文自述为"单一奇幻酒馆场景的实践导向比较案例研究"，**不能推广为"AI 场景的普遍属性"**；它也不能说明"AI 一定更慢"。它说明的是一件更朴素的事：**把生成物的默认输出直接当成生产物，会在下游付出代价。**

图 7-11　111 倍的面数差距从哪来。注意第二行采用对数刻度，否则两根对比条的长度比例会让人误以为差距只有五倍左右。数据来源：三维模型生成研究报告中引用的 n=1 对照案例。

🩺 **失败诊所 F7-2** 风格混杂与比例失调："拼盘场景"

症状：场景里每一件单独看都不错，合起来却像一盘拼盘：楼台是写实的、扁舟是低多边形的、远山是水墨的、灯笼是卡通渲染的；或者更隐蔽的一种——所有东西风格一致，但**比例全乱**：门高 6 米、船宽 12 米、栏杆比人还细，镜头一推近就露馅。

成因：两个独立的原因叠加。**原因一：多次生成之间没有共享"约束"。**每次调用生成工具都是一次独立抽签，如果你每次都只靠文字描述，风格就一定会漂移——这正是"级联误差"的典型场景：每一步偏一点，最后整体跑偏。**原因二：比例从来没有被显式声明过。**AI 模型常以"任意真实世界尺度"导出（表 7-7 的第一条坑）；当"楼台"的生成结果高度是 9 单位、"扁舟"是 3.2 单位而它们各自的单位含义不同时，拼在一起就是灾难。**换句话说：风格漂移是"参考不统一"造成的，比例失调是"单位不统一"造成的——两者都不是审美问题。**

处方：三条，成本都很低。**① 建参考图集与色板**，全组共用一套，生成任何资产前先看一遍；能用风格包/预设的地方一定用预设（这也是天空盒工具提供 30 余种风格包的现实意义——它把"风格统一"变成了一个下拉框）。**② 在场景里放一个"参照物"**：一个 1.8 米高的人形盒或一根标尺柱，任何资产导入后先与它比一次高度；这一招能当场抓住九成的比例问题。**③ 固化单位与变换**：约定 1 单位 = 1 米，导出前应用变换，让缩放读数变成 (1, 1, 1)——这不只是工程规范的洁癖，它是"下一个导入这个文件的人能不能用"的分界线。

一句话诊断法：当你说不出"这个场景哪里不对"时，先做两件事——把它全部换成纯灰白材质看一次（**风格伪装被剥掉后，比例问题会立刻显形**），再把它截图缩小到手机屏幕大小看一次（**细节的嘈杂消失后，风格冲突会立刻显形**）。

🩺 **失败诊所 F7-3** 符号误用：当"夏季西瓜"和"春季樱花"同框

症状：画面很美，但懂行的人一眼看出不对——同一场景里出现了"夏季的西瓜"和"春季的樱花"；建筑挂着某个朝代的形制、甲胄却是另一个朝代的风格；某个族群的角色穿着另一个族群的服饰。

成因：这是"符号"与"事实"的混淆。生成模型学的是**符号的共现规律**——"东方""古代""庭院"这些词在训练数据里经常和某些元素一起出现，于是它会把这些元素打包输出。但它不区分这些元素的**年代、地域与使用场合**。检索到的争议案例中，最典型的两个：其一，某款以日本战国为背景的作品在 2025 年被报道出现"夏季西瓜与春季樱花同框"的**季节矛盾**，并有学者批评角色服装形制与其设定身份不符；其二，某款历史策略游戏的限定内容中，城市国家的角色穿着被指为另一族群的服饰，争议归因包括"同文化圈模型复用""默认服饰复用"以及长期考据不足——在第三次补丁后，该内容才被替换为专门的服饰。

处方：不是"少用东方元素"，而是**把考据变成流程里的一步**。检索到的正面路径给出了可操作的方法论：某款历史策略游戏在 2026 年 3 月推出的一套由中国创作者主导的服饰内容包，其考据方式是**参考文物与古画**——例如某种袍服参考了具体的明代行乐图卷与墓葬出土实物，另一种袍服参考了辽代墓葬出土的左衽交领袍。报道认为这种做法"减少了文化挪用与想当然还原的担忧"。把这条经验移植到你的作业里，就是三句可执行的话：**一、凡涉及历史文化元素，先找一件可追溯的实物或古画参照，把出处写进资产条目的备注字段；二、找不到可靠参照时，主动把设计"抽象化"（用意象而非具体形制），并在作品说明里如实写明这是风格化处理；三、宁可少用，不可错用。**

为什么这条对通识课尤其重要：因为最容易犯这个错的，恰恰是"有热情但没考据"的创作者。下一节的思政专节会把这条展开——它不只是审美问题，还牵涉到你作品的传播后果。

7.10 动手实践：搭出你自己的"夜江场景"

本节共四个实践，前三个课堂完成、第四个课后完成，合计约 3 学时（其中实践二可与课后时间合并）。全部实践围绕同一个产出物：**一份场景初稿 + 一份结构化资产清单**——这既是本章的主线推进，也是课程作业"提交场景初稿与结构化资产清单"的内容。

先给你一张本章的"施工总图"。左侧是俯视布局（示意，非等比），右侧是它对应的清单——**清单和布局是同一件事的两种表达**：前者给机器读，后者给人看。

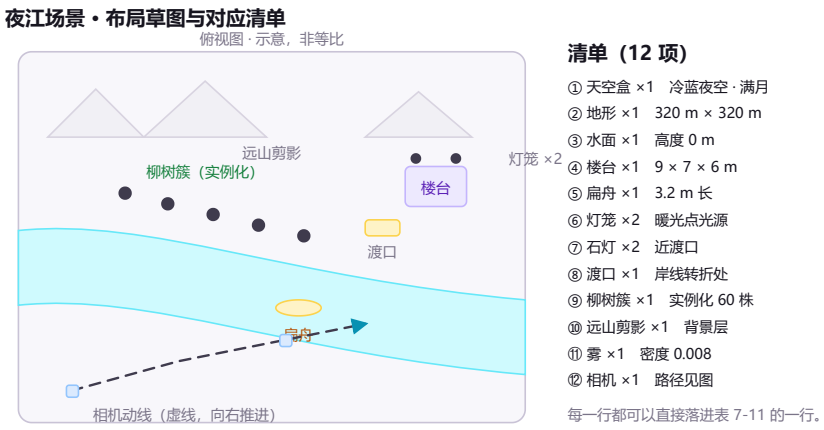


图 7-12　夜江场景的布局草图与对应清单。先有清单、再有布局，或者先有布局草图、再补出清单——两种顺序都可以，但两者必须能一一对应。任一侧有孤立的东西（清单里有而图上没有，或图上有而清单里没有），就是返工的来源。

图 7-13　主线场景的目标态（夜江场景概念图）。在写资产清单之前，先有一张「目标态」。把它拆成清单的过程正是 7.3 节的全部内容：前景石岸与花树、中景江面与小舟、远景群山与楼台灯火——每一块都对应若干条资产需求。 AI 生成

🛠 实践一：场景方案设计——让模型帮你写资产清单（40 分钟 • 课堂）

基础

四步走，产出你的 `scene.json` 与一页布局草图：

1. **写需求。**用 100 字以内描述你想做的场景（可以是夜江，也可以换成你喜欢的任何题材，例如"雨夜的旧巷""雪山下的营地"）。**必须点名**：时刻、天气、主体建筑、一件核心道具、情绪词。
2. **出清单。**把 7.3 节的"生成场景规格表"提示词卡复制进对话助手，把方括号里的内容换成你的需求；把表 7-4 的字段清单粘进提示词。要求模型只输出一个 JSON 对象。
3. **做四道校验。**对照 7.3 节结尾的四道关：schema 是否合规（字段名和类型对得上吗）、数量是否超上限、相机路径是否走得通、面数预算的加总是否自洽。把不合规的地方**改回 JSON**，不要只在心里记住。
4. **画布局草图。**照图 7-12 的样子，用纸笔或绘图工具画一张俯视图，标出清单里每一件的编号。凡是在图上找不到位置的资产，就说明它其实不该出现在清单里。

提交物： `scene.json` + 一张标注了编号的布局草图。评分看两条：清单能否被脚本解析；清单与草图是否一一对应。

挑战

给模型增加一个"目录约束"：限定资产条目不超过 12 条、建筑不超过 2 座、同类道具不超过 3 件，并要求它在超限时**自己说明舍弃了什么、为什么**。你会看到"约束"如何把一份散漫的清单变成一份有取舍的方案——这正是 AI PCG-2D 一类项目里"硬上限"设计的用意。

🛠 实践二：资产生成——环境层与物体层（80 分钟 • 课堂 + 课后）

基础

环境层（先做，约 30 分钟）：

1. **天空盒。**用 7.5 节的天空盒提示词卡，写清时刻、光照方向、云量、色调，并明确"无缝衔接""无人物、无动物、无文字"。导出后做 360° 环视，检查接缝与拉伸。
2. **地形。**按 7.4 节的三个旋钮（层数 / 频率增量 / 振幅衰减）生成高度图，参数与种子写进 `scene.json`；再手工把楼台落位处压平、把江岸修出一条缓坡。
3. **水面。**用平面 + 水体材质；确认反射来源指向天空盒、水面高度与地形共用坐标系。

基础

物体层（约 50 分钟）：用下面的提示词卡分别生成建筑与道具，每类至少生成 2 版、择优采用。

✎ 提示词 • 建筑 • 临江楼台

[唐代]风格的临江楼台，两层，木构架、歇山顶、出檐深远，带[栏杆与窗棂]；体量约 9 米宽、7 米高、6 米深；[朴素含蓄，不施彩画]；参考图为下图；纯几何、无环境、无人物；输出 GLB

参考图 必传 1–2 张　面数 ≤ 8000　LOD 3 级　Pivot 地平面中心

✎ 提示词 • 道具 • 扁舟

[一叶小舟]，木质，长约 3.2 米、宽约 1.1 米，船身微翘、带一支短桨；[古朴、无装饰]；材质偏粗木，颜色[深褐]；纯几何、无环境、无人物；输出 GLB

面数 ≤ 1200　Pivot 几何中心　贴图 2K 以内

进阶

植被与材质：用泊松盘采样（或抖动网格）沿江岸散布 40–60 株低模柳树，并在引擎中**以实例方式**复用同一个模型；材质统一压到 2K 以内，数据贴图关闭 sRGB 色彩空间，粗糙度贴图若来自外部工具记得在 Unity 里取反。

🛠 实践三：拼装与检查——把八项检查表跑一遍（40 分钟 • 课堂）

基础

按图 7-9 的八项检查逐条走，边查边填表 7-12 的记录。三条硬要求：**① 场面里必须放一个 1.8 米高的人形参照物**；**② 每件资产导入后都要看一次统计面板的面数**；**③ 发现的问题必须记录"改动前 / 改动后"**，这正是表 7-13 的任务。

进阶

做两次"对照自检"：**A 组**只把资产拖进引擎、肉眼看着差不多就过；**B 组**完整跑八项检查。两组各记录：发现的问题数、修掉的问题数、花掉的时间。把两组数据填进下面的对照表——你会发现"检查"花掉的时间，正是从"反复返工"里省下来的。

A 组 • 只靠肉眼

导入快，第一印象好；问题在演示时才暴露：楼台比远山大、船横躺在草地上、贴图是粉的。

代价：发现问题的时刻最晚，修起来最贵。

B 组 • 跑完八项检查

导入慢一点，但问题当场记录、当场修；交付前不需再"救火"。

收益：问题修在"改一行数字"的阶段，而不是"推翻重做"的阶段。

🛠 实践四：性能初检与优化（30 分钟 • 课后）

进阶

1. **读数字。**打开引擎的场景统计面板，记录三角面总数、贴图内存占用、独立材质数量。三项都要写进作业文档，不能只写"感觉很流畅"。
2. **对照预算。**与 `scene.json` 里的 `budget` 比对；超了就先做三件事：给远处的树和建筑生成 LOD、把重复资产改成实例、把 4K 贴图降到 2K。
3. **再读一次数字。**把优化前后的两组数字并排列出（表 7-13 的下半部分），并写出你的结论一句话。
4. **记录未解决的问题。**诚实写下"我知道它有问题但这次没解决"的那几项——比如某件道具的 UV 还没重做。**把技术债写出来，是让它不再复利增长的第一步。**

提交物：优化前后对照表 + 一句结论 + 未解决问题清单。

挑战

把你的场景清单交给同组另一位同学，让他**只按清单**（不看你的场景）复述"这个场景长什么样"。如果他的复述和你脑中的画面有明显偏差，说明你的清单还缺字段——通常是缺了空间关系（谁在前、谁在后）或尺度参照。

图 7-14　四个实践的完整操作路径。从左到右是一条单行道：清单定约束、环境定光线、物体提供内容、整合负责验收。中途"回头改清单"完全可以，但要记得把改动同步回 `scene.json`，否则清单就成了一份过期的说明书。

表 7-11　场景资产清单模板（以夜江场景为例，可复制为你的空表）

编号	类型	名称	生成方式	尺寸 (m)	位置 (x, y, z)	面数上限	LOD	状态与备注
sky_001	天空盒	夜江夜空	文生 360°	–	–	–	–	已生成 · HDR 导出
terr_001	地形	江岸丘陵	fBm 噪声 + 雕刻	320 × 320	0, 0, 0	40,000	2	已生成 · 种子 20261009
water_001	水面	夜江江面	平面 + 水体材质	320 × 320	0, 0, 0	200	1	已生成 · 高度与地形一致
bld_001	建筑	临江楼台	文生三维 + 参考图	9 × 7 × 6	−6, 0, 22	8,000	3	待重做 · 形制需核实
prop_001	道具	扁舟	文生三维	3.2 × 0.6 × 1.1	4, 0, 30	1,200	2	已生成 · 一次通过
prop_002	道具	灯笼 ×2	文生三维	0.4 × 0.5 × 0.4	−3, 2.4, 23	400	–	已生成 · 暖光点光源载体
veg_001	植被	江岸柳树簇	低模 + 泊松盘散布	高 4–6	沿江岸	600 / 株	3	已生成 · 实例化 60 株

备注栏必须写：考据出处（文物 / 古画 / 参考文献）、授权状态、生成工具与日期——这三项决定了你的作品能不能被复用。

表 7-12　场景初稿的前后对照（作业必填）

项目	修改前	修改后	触发修改的检查项
楼台面数	约 12 万面（工具默认导出）	7,800 面	预算自洽校验（面数上限）
楼台贴图	4K 基础色 + 4K 法线	2K + 1K	贴图内存检查
整体比例	各资产"看起来差不多"	与人形参照盒逐件比对	检查表第 1 项（尺度与单位）
Pivot	位于世界原点	移至地平面中心	检查表第 4 项（Pivot 与原点）
柳树	60 个独立模型	1 个模型 + 60 个实例	绘制调用与内存优化
水面反射	反射方向与月亮不一致	材质反射源指向天空盒	视觉统一检查
相机动线	路径穿过楼台内部	路径外移 4 米绕行	玩法规则校验（连通性）

表 7-13　导出前自检记录表（实践三填写）

序号	检查项	判定方式（读数字，不看图）	通过	备注 / 待修
1	尺度与单位	与人形参照盒（1.8 m）比对高度	□	
2	变换是否已应用	缩放读数是否为 (1, 1, 1)	□	
3	轴向与朝向	导入后是否直立、是否对齐前向轴	□	
4	Pivot 与原点	环境资产在地平面、道具在几何中心	□	
5	面数与拓扑	统计面板面数是否落在预算内；非流形边是否为零	□	
6	UV 展开	是否有重叠壳、缺失 UV；纹素密度是否一致	□	
7	材质与贴图通道	粗糙度是否需反转；数据贴图是否关闭 sRGB	□	
8	碰撞体与 LOD	是否已生成简化碰撞体；远景是否有 LOD 链	□	

7.11 课程思政：文化准确性与团队劳动

这一节讨论两个容易被回避、却每天都发生在场景创作里的问题：**你搭的这个"世界"，对它所代表的文化负责吗？搭它的这些人（包括你自己）的劳动，被诚实对待了吗？**我们一条一条说清楚，不给空话。

◆ 一、场景是文化载体：准确比"好看"更重要

前面讲过，建筑的第二个职责是"提供文化信息"——观众看到飞檐斗拱、看到服饰形制，就会自动推断"这是哪里、什么年代、谁住在这里"。这意味着**场景设计天然是一次文化陈述**，它不可能"只是美术"。

检索到的公开争议提供了几条真实教训。一款以日本战国为背景的作品在 2025 年被报道出现"夏季西瓜与春季樱花同框"的**季节矛盾**，并有学者指出角色服装的形制与其设定身份不符——争议的核心不是"画得美不美"，而是"考据粗疏"。同一批报道还提到：该作品中某处宗教场所可以被破坏，相关神社与协会提出抗议，据报道未获回复、补丁后仍可破坏，事件一度进入该国国会讨论层面。**这类后果远超"美术问题"：当场景把一个真实的信仰场所处理成可随意破坏的对象时，它触碰的是许多人的真实感情。**

另有一款以敦煌为背景的作品被批评建筑风格"过度中亚化"，专家认为应更具唐风，理由很具体：高昌出土的唐代木构是典型唐制，晚唐时期唐文化仍是该区域的主流文化。也有反驳意见认为，官式与宗教建筑可能更唐制，而普通民宅未必如此——**这个反驳本身很有价值：它提醒我们，"考据"不等于"想当然地统一化"，不同阶层、不同功能的建筑本来就有差异。**还有一款历史策略游戏在限定内容中，让某座城市国家的角色穿上了另一族群的服饰，争议归因包括"同文化圈模型复用""默认服饰复用"以及长期考据不足；直到第三次补丁，才替换为专门的服饰。

在讲了这些争议之后，必须补上一句平衡的话：网络舆论里也出现了不同国家、不同地区的网友互相指责"挪用"与"双标"的现象——**本书不站队，也不认为存在一个"谁有资格做谁的文化"的简单答案。**真正有共识的，是**方法论层面**的那一条：

> *可追溯的考据来源 > 风格化想象*
>
> 这是全书关于文化准确性的基本立场：不限制你创作什么，但要求你为每一个具体的形制找出处

正面路径是存在的，而且很具体：某款历史策略游戏在 2026 年 3 月推出了三款创作者内容包，涵盖东亚、东北亚与东南亚的服饰与建筑；其中一套是该系列**首次由中国创作者主导、官方发行**的内容包，作者来自上海。它的考据方法可以直接被我们借鉴：**参考文物与古画**——例如某类袍服参考了《宣宗行乐图》《明宪宗元宵行乐图》等画作与明代墓葬出土实物；另一种袍服参考了辽代墓葬出土的左衽交领袍。报道的评价是：过去海外游戏处理东方题材常"元素堆得不少，形制却经不起推敲"，而由本国创作者操刀、依据文物古画考据的做法，减少了"文化挪用"与"想当然还原"的担忧。

> **🇨🇳 课程思政 · 用场景讲好中国故事：三条可操作的规矩**
>
> 把上面的教训落成你能执行的动作，只有三条：
>
> **① 出处优先。**凡涉及历史文化元素的资产（建筑形制、服饰、器物、节令物候），在清单的备注栏写下它的考据来源：是某幅古画、某处出土实物，还是某份文献。写不出出处的，就不要用它当"历史元素"，而应明确标注为"虚构/风格化"。**本章表 7-11 的备注栏就是为此而设。**
>
> **② 少用优于错用。**一个"意象化"的江边剪影，胜过一个形制错误的唐楼。用意象、光线、留白去表达时代感，比用一组错误的符号更安全，也更高级——这也是水墨风格在传统文化题材里特别适合初学者上手的原因。
>
> **③ 说清楚你在做什么。**在作品说明里写明白："本场景为艺术创作的风格化演绎，建筑形制参考某时代某类建筑，非严格复原。"这句话不削弱作品，反而体现了创作者的自觉——它把"风格化"与"考据"两种不同的标准分开了。

◆ 二、团队劳动：把"效率"和"代价"一起算清楚

第二个话题更贴近日常。这一章反复出现了一组数字：**面数差 111 倍、建模省下 478 分钟、这些时间转化成了返工的技术债。**请再读一遍最后半句。

行业公开数据里，效率数字往往只算"生成"这一段：三维资产生产从"5–7 天、数千元"到"10 分钟、数元"，听上去像魔法。但同一批检索材料也给出了另一半：普通的手动重拓扑，中等道具约需 30–60 分钟、角色约需 2–4 小时；AI 生成的网格常常"带有数十万个近似均匀的三角面，对渲染、编辑和动画管线而言过于密集"，而纯粹的几何减面"不尊重尖锐特征与部件结构"，达不到生产资产要求的干净拓扑。**把这两半合起来看，结论就很清楚：AI 没有消灭这些劳动，它把劳动的位置从"从零开始做"移动到了"检查、判断与重做"。**

这个移动有三层含义，都值得你在课程里认真体会：

第一层，谁的工作变了。在盛趣的那个案例里，一个局部场景复刻投入的是"1 名三维艺术家 + 1 位场景设计师 + 半个技术支持"。没有一个是"专职点生成按钮"的岗位——因为按钮太便宜了。人力向三处集中：**定义需求、甄别资产、把检查自动化。**这和本章 7.3 节讲的"结构层交给脚本、创作层交给人"是同一件事的两面。

第二层，谁的判断不可替代。模型能生成一万座"楼台"，但没有一座能替你回答"这座楼台放在这个镜头里对不对、它的形制对不对、它与小月定的色板和不和"。在本章的作业里，这三问恰好对应三个不同的岗位视角：小舟管"对不对"（技术检查），小月管"和不和"（视觉统一），小江管"该不该有"（叙事必要）。**小组分工的价值，不在于每人分到多少活，而在于每人都保留了一块"必须自己判断"的地盘。**

第三层，怎么诚实地写效率。检索材料里有一个反复出现的提醒：行业里大量效率提升数据来自**企业自称或券商测算，口径各异**，不能当作客观测量结果。本书在表 7-10 中为每一条数据都标注了来源性质——这不是走形式，这是通识课要教给你的媒介素养：**看到"提效 300%"时，先问三句——口径是什么？分母是什么？算没算返工？**同理，在你自己的作业报告里，也请诚实写下"哪部分我用了 AI、哪部分我手工做了、返工花了多久"。一份诚实的报告，比一份漂亮的报告有价值得多。

🇨🇳 课程思政 · 尊重劳动：三件小事，映射同一个态度

其一，如实标注 AI 参与。发布作品时按第 1 章讲过的要求做 AI 生成内容的标识——这既是法规要求（《人工智能生成合成内容标识办法》自 2025 年 9 月 1 日施行），也是一个创作者对受众的基本诚实。技术说明与参数对照要写清楚：哪些资产是生成的、用了什么工具、做了哪些人工修改。

其二，尊重素材来源与授权。参考图、扫描数据、动作数据都可能带着使用条款；把它们记进资产清单的备注栏（来源、授权方式、下载日期），是对他人劳动的尊重，也是作品能否被你继续使用的前提。

其三，别把"快"当成唯一标准。这一章最想留下的一句话，来自小舟写在小组日志里的那一句："省下来的不是做资产的时间，而是反复返工的时间。"**AI 给的是速度，不是判断；是产能，不是审美；是更多个选项，不是正确答案。**把这三句话记住，你就已经比大多数"会用工具的人"走得更远了。

◆ 把这一章压缩成七句话

1. 场景由**地形、建筑、道具、光照、氛围**五块拼图构成：地形给"地"，建筑给"形"，道具给"事"，光照给"真"，氛围给"情"；拍照不对时，先问自己缺的是哪一块。
2. 场景"好不好用"要用三个维度验收：**视觉统一、玩法动线、性能预算**——三者互相拉扯，先决定谁优先、谁让步，再动手。
3. 大语言模型在场景任务里最可靠的用法是"**翻译**"：把自然语言需求转成结构化的 JSON 规格表，而且必须"强制结构化输出 + schema 校验 + 玩法规则校验"，**校验先于实例化。**
4. 结构层（严格 JSON、可自动校验、失败可重试）与创作层（自由散文、由人审阅）必须分开——前者交给脚本，后者交给你自己。
5. 程序化生成的核心是"**在规则允许的范围内随机**"：噪声管地形，波函数坍缩管受约束的布局，泊松盘管自然散布，种子的意义是"同一个输入必然得到同一片地形"。
6. 引擎落地的四个不可见的坑是**尺度、变换、轴向、Pivot**；验证必须在导出前完成，用确定性断言读数字而不是看截图——"引擎导入决定资产能否出货"。
7. 最后一条是本章的价值观：把生成的默认输出直接当成生产物，代价会以**返工与重做**的形式出现；所以既要尊重文化的考据来源，也要诚实计算团队劳动的成本。

◆ 自测

❓ 自测 1：一位同学说："我让 AI 生成了一个场景，拖进引擎就能用。"请指出这句话里至少两个需要怀疑的地方。

答案：至少有两处。① **"生成"不等于"生产可用"**——生成器的默认输出在面数、拓扑、UV、碰撞体上通常不满足项目预算，本章的 n=1 对照案例里，AI 辅助场景的面数是手工场景的约 111 倍，省下的 478 分钟建模时间变成了返工技术债。② **"拖进引擎就能用"忽略了导入的四个坑**——尺度（米与厘米差 100 倍）、未应用的变换（到引擎里会变成"煎饼"）、轴向（可能"横着到达"）、Pivot 偏离（放置成为噩梦）。正确的说法是："它看起来能用，但要先过一遍图 7-9 的八项检查，才知道它是不是真的能用。"

❓ 自测 2：为什么本书要求"先校验、再实例化"，而不是"先生成出来看看再说"？

答案：因为两件事的**纠错成本差了几个数量级。**规格表里的错误，纠正它只要改一行数字或重新生成一次；而实例化之后的错误——比如相机动线穿过楼台、资产数量超标、预算超支——往往意味着已经摆好的东西要重新排布、已经生成并导入的资产要重做。检索到的可运行项目几乎都遵循这条纪律：对 JSON 做严格连通性校验（确保出生点到出口存在通路）、做目录约束（最大对象上限）之后再建场景；其中有的项目更进一步，坚持"先设计解法路径、再写地图 JSON"。生活类比：装修要先验线，而不是等墙砌完再砸。

❓ 自测 3：用"三个旋钮"解释：为什么单层随机噪声看起来不像地形，而分形噪声（fBm）像？

答案：因为真实地形同时具有"大尺度的起伏"和"小尺度的粗糙"两种特征，而单层噪声只能提供其中一种：频率低、振幅大的那一层给你宽阔的山形，但表面太光滑；频率高、振幅小的那一层给你粗糙的质感，但整体没有高低。分形噪声把多层不同频率的噪声叠起来，三个旋钮分别控制叠几层（octave）、每层频率乘几倍（lacunarity，常用 2.0）、每层振幅乘几倍（persistence，常用 0.5），于是就同时得到了"山形"与"质感"。另外要记住：噪声是伪随机的——相同输入得到相同输出，因此固定种子就能复现同一片地形；而在性能上，绝不能每帧采样噪声，必须预生成一次高度图再从内存读取。

❓ 自测 4：波函数坍缩（WFC）为什么要"先挑熵最低的格子"？如果先挑熵最高的会怎样？

答案：熵在这里的具体含义是"这个格子的候选图案数量"。先挑熵最低的格子，等于**先做最有把握的决定**——候选最少意味着约束最多，一旦确定，就能可靠地把大量不兼容的图案从邻格候选里删掉，让后续每一步都更容易。反过来，先挑熵最高的格子等于在最不确定的地方乱猜：一旦猜错，错误会沿着传播链扩散，最终使某个格子的候选被删空，产生"矛盾"，只能回溯或重启（这也是 WFC 必须设最大迭代次数安全阀的原因）。顺带记住那条实用建议：**简单布局不要用复杂 WFC**，改用醉汉漫步或 BSP 更合适。

❓ 自测 5：有同学说："我的场景在引擎里看着没问题，说明资产是合格的。"请用本章的工程观点反驳，并给出两种替代的检查方式。

答案：反驳的依据是那条工程结论：截图或精美的转台**只证明了外观**，它不证明尺度、不证明变换是否已应用、不证明 Pivot 是否合理、不证明有没有非流形边、也不证明面数是否在预算内——因为这些属性只有"在特定上下文里"才会显形。替代方式（任答两种）：① **确定性断言**，例如检查缩放读数是否为 (1, 1, 1)、非流形边计数是否为零；② **导出回读**，导出后重新读回文件，验证数据与导出前一致；③ **读统计面板**，用面数、贴图内存、独立材质数量这些数字与规格表里的预算逐项比对；④ **放参照物**，用 1.8 米的人形盒比对每件资产的高度。核心原则一句话："验证必须在导出前完成——引擎导入是最终检查，不是调试阶段。"

❓ 自测 6：某作品被批评"把唐代建筑画成了和风样式"。请分别从"技术流程"和"创作伦理"两个角度说明你该如何改进。

答案：技术流程角度：问题出在流程里没有"考据"这一环——只靠文字提示词生成，模型会把"东方古代建筑"的符号打包输出，而它并不区分年代与地域。改进办法是把考据变成流程的一步：先从文物、古画、考古报告中确定形制（例如某类木构的具体做法），把来源写进资产清单的备注字段；生成时提供参考图而不只是文字；把"形制核对"列为验收项之一（与面数、UV 同级）。**创作伦理角度：**要认识到场景是文化陈述，错误形制可能造成误读与冒犯，尤其是涉及宗教场所、历史疆域等敏感内容时后果更严重；态度上采取三条规矩——出处优先（写不出出处就不当作历史元素使用）、少用优于错用（用意象与光线表达时代感）、说清楚你在做什么（在作品说明中标明是风格化演绎还是严格复原）。这里的方法论共识可以概括为一句：**可追溯的考据来源 > 风格化想象。**

◆ 拓展阅读

- **随查手册：**本章用到的三维、天空盒、智能体平台等工具选型见附录 B 工具速查；场景资产涉及文化元素与真人素材时的合规要点见附录 E；把资产生成流程自动化的路径见附录 F。
- **Skybox AI（Blockade Labs）官方文档与"2026 完整指南"**——天空盒生成的完整工作流、风格包、导出格式与免费档限制；建议亲手生成 3–5 个不同时刻（清晨/正午/黄昏/夜晚）的同一场景天空盒，比较它们对同一批资产观感的影响（本章 7.5 节）。
- **Gumin 的波函数坍缩原始项目与讲解文章（2016）**——如果你想真正理解"观察—传播"循环，最有效的方式是找一个网页版 WFC 演示，亲手调小网格与样例图，观察它什么时候会失败、什么时候需要重启（本章 7.4 节）。
- **引擎官方文档中的"资产导入与优化"章节（Unity / Unreal / Godot 均有中文页面）**——重点读"缩放因子、轴向、LOD、光照贴图"四节；把本章图 7-9 的八项检查与官方文档逐条对照，形成你自己的检查表（本章 7.6 节）。
- **与"场景 AI 生产"相关的行业报告与公开案例（伽马数据、中国音数协、各厂商公开披露）**——阅读时请刻意练习一件事：为每一个效率数字标注"来源性质"（企业自称 / 券商测算 / 第三方测量），并追问它的口径与分母（本章 7.8、7.11 节）。
- **文博机构公开的书画与器物高清资源（如各大博物馆的数字藏品页）**——这是最便宜也最可靠的"考据参考图集"来源；练习用它作为生成建筑与器物资产时的参考图，并在清单备注里写下具体条目（本章 7.11 节）。

第 8 章

智能体与交互媒体创作

从"会聊天"到"会办事"：工具、记忆、规划与验收，以及一套真正把前七章串起来的工作流。

学习目标

1. **知识：**理解智能体的四要素（大语言模型 + 工具 + 记忆 + 规划）；能说明 ReAct 循环、工具调用（Function Calling）与模型上下文协议（MCP）的作用；理解检索增强生成（RAG）的流程与常见失败；知道多智能体协作的三种拓扑及其代价。
2. **能力：**能在零代码平台上搭建一个可用的智能体工作流；能把"完成一部短片"分解为可验收的子任务；能设计并执行"单智能体 vs 多智能体"的对照实验；能用验收清单对 AI 产出做分档判定。
3. **素质：**建立"人在回路"的责任意识，明确人与智能体的分工；坚持创作主体性与透明协作，如实标注 AI 参与程度、如实说明个人贡献，杜绝挂名与抄袭。

对应第 10–11 周 · 5 学时　讲授 3 学时 · 实践 2 学时　主线推进：多智能体合成与期末展示验收

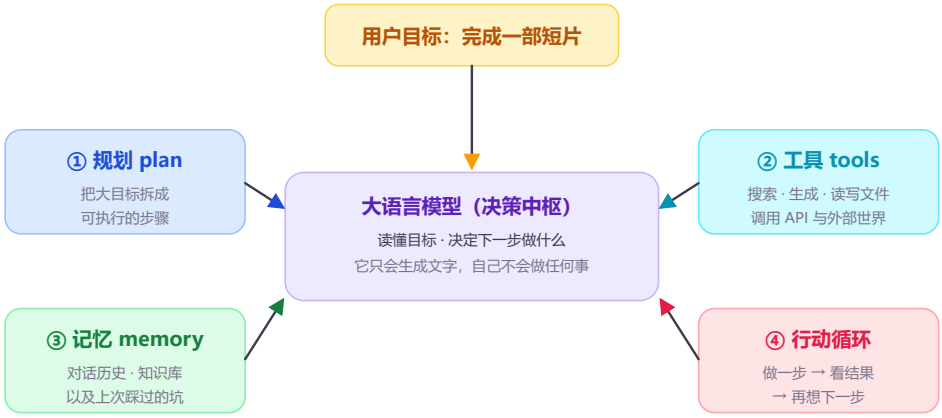
> 期末前两周，小舟打开那个装着全部素材的文件夹：小江定稿的脚本与配音、小月做的分镜图、第 3 章的镜头视频、第 4 章的三维资产、第 5、6 章的动画与动作、第 7 章的夜江场景——七个环节、几十个文件、十几份参数记录，全都散落着。**它们都合格，但它们是"零件"，还不是"作品"。**小舟打开对话框，输入："帮我把这部短片串起来。"对方回了一段漂亮的建议清单：先做粗剪、再统一色调、然后合成音轨、最后导出——条理清晰，面面俱到，然后……就没有然后了。清单不会自己执行，它只是又一页待办事项。

这一幕点破了本章要讲的核心差别：聊天助手给你"方案"，智能体（agent）替你"办事"。把"方案"变成"动作"的能力——调用工具、记住上下文、规划步骤、检查结果——正是生成式 AI 从 2024 年的"对话框"走到 2026 年的"工作流"的关键一跃。本章要回答三个问题：**第一，**智能体凭什么能从"聊天"升级为"办事"，它的四个零件各自管什么？**第二，**这些零件怎么装到一起——工具调用、模型上下文协议、检索增强生成与记忆，各自解决什么问题、又会在哪里失败？**第三，**一个更反直觉、也更有价值的问题：多智能体一定比单智能体强吗？如果答案是否定的，我们该怎么算这笔账？

这一章也是全书的收官章。学完它，你不仅能给春江小组的《春江花月夜》装上一套自动流水线，还能拿到一套可以迁移到任何项目的"方法与清单"——以及一个关于"人在哪里"的、诚实的结论。

8.1 智能体：四个零件，一个"会办事"的系统

先给一个定义：**智能体**（agent）是一个以大语言模型为决策核心、能够感知环境、调用工具、保存记忆并在循环中推进任务的系统。2026 年的多份综述给了它一个高度一致的四要素分解——大语言模型 + 工具 + 记忆 + 规划。我们把这四个零件先摆开看，再讲它们怎么转起来。



四个方框缺一不可：没有工具 = 只会说；没有记忆 = 每轮从零开始；
没有规划 = 走一步看一步；没有循环 = 做完不检查、错了不回头。
本章要做的，就是给前七章的资产配上一套"会办事"的流水线。

图 8-1　智能体的四要素。中间是"只会生成文字"的大语言模型；四边是让它能办事的四件装备。聊天机器人只有中间一格，智能体把四格配齐并让它们循环运转。

一个类比：从"会聊天的访客"到"有工位的实习生"

想象你带过两种新人。第一种是**访客**：你问他"这片子怎么剪"，他能说得头头是道，但你让他"去把第三段的音轨拉齐"，他只能说"好的，建议先打开剪辑软件，然后……"——**他能说，不能做**。第二种是**有工位的实习生**：他有门禁卡（能进机房取素材）、有软件账号（能导出文件）、有一本工作笔记（记着你上周交代的风格要求）、还有一张任务清单（知道今天该交付什么）。你交代一句"把第三段音轨拉齐"，他去做了，做完把结果放在你桌上。

四件装备正是一一对应：**工具** = 门禁卡和账号（能把动作施加到外部世界）；**记忆** = 工作笔记（不必每次都从零说起）；**规划** = 任务清单（知道先做哪一步、后做哪一步）；**行动循环** = 做完回来看一眼结果、判断要不要重做。这套类比里最要紧的一句是：大语言模型本身，永远是那个"只会说话"的部分。智能体的全部本事，来自围绕它搭起来的这套"工位"。

表 8-1　聊天助手与智能体：七个维度对照

对比项	聊天助手	智能体
你给它的	一句话、一段材料	一个**目标**（如"把分镜 3 改成夜景"）
它还给你的	一段文字	一个**做完的结果**（文件、链接、订单、发布的内容）
能不能"动手"	不能，只能生成文字	能调用工具：搜索、生成、读写文件、调用外部服务
记不记得住	单次会话内；换对话框即失忆	有外部记忆：对话历史、知识库、跨会话的长期记忆
谁来推进	你——每一步都要你再发一次指令	它在循环里自己推进，直到完成或明确放弃
典型失败	说错话（幻觉、跑题）	做错事（误操作、原地绕圈、偏离目标、越权）
你的角色	提问者	监督者与验收人——**责任更重，而不是更轻**

表 8-1 的最后一行值得单独停一下。从聊天到办事，人的位置不是"退场"，而是从"每一步的操作员"变成"节点上的验收官"。这也是本章最后一节要反复回到的立场：智能体放大的不是你的懒惰，而是你的判断力与责任感。

四件装备，缺一件会怎样

概念讲完，来看故障。教学上最有效的方式是把四要素逐个"摘掉"，观察系统退化成什么样子——这也正是你调试自己搭的第一个智能体时最可能遇到的四种情形。

表 8-2　缺一件装备时，智能体退化成什么样

缺哪一件	表现症状	排查方向
没有工具	输出一堆"建议你先……然后……"，看起来很专业，但没有产生任何文件或动作	平台里是否挂了插件/工具？工作流里是否有"执行"类节点？
没有记忆	第二轮又在问你"请问您想要什么风格"，或反复重做已经确认过的事	是否开启了长期记忆？是否把关键结论写进了提示词或知识库？
没有规划	开头几步还行，做着做着跑到别处去了；或者把一个大任务草草"一次性"交差	目标是否写清楚？是否要求它先列步骤、再逐步执行？
没有循环与反馈	一口气输出结果，从不查看工具返回；失败也不重试，直接把错误当结果交给你	是否给了它"看结果"的机会？有没有设定失败后的重试与上报策略？

把这四种症状和表 8-2 的右列对照着记，你在后面实操时的调试速度会快很多：**看到症状，先定位是哪件装备缺了，再去平台里找对应的开关**——而不是笼统地"把提示词再写得客气一点"。

> 🧪 **原理注 · 智能体不是新模型，而是新用法**
>
> 一个常见误解是"智能体是更高级的模型"。不是。同一个大语言模型，被你直接用来一问一答时是"聊天助手"；被套上工具、记忆、规划的脚手架、放进循环里时，就成了"智能体"。**智能体的能力上限由模型决定，但它能做到的事，主要由脚手架决定**——这也解释了为什么 2025 年以来"智能体"的进步有相当一部分不是模型变强，而是工程变好（工具设计更规范、上下文管理更精细、停止条件更可靠）。反过来说：你在这门课里学到的很多东西——任务分解、验收标准、上下文组织——恰恰是那个"脚手架"的图纸，它们不会随模型换代而过时。

8.2 ReAct：想一步、做一步、看结果

智能体最经典、也是采用率最高的运转方式叫 **ReAct**——这个名字由"推理"（Reasoning）与"行动"（Acting）两个词拼成。它规定了一个朴素的循环：想一步（Thought）→ 做一步（Action）→ 看结果（Observation）→ 再想。就这样简单。

为什么这样一个循环如此有效？用"找东西"来体会：让你找一个不知道放在哪儿的文件，你并不会"一次性想清楚它的位置"，而是——打开第一个抽屉看一眼（做），没找到（看结果），换个思路去翻文件夹（再想），找到了（结束）。**真实世界的任务，本来就是"边做边看"的。**传统的一问一答把模型锁在"只能想、不能看"的位置上；ReAct 松开这只手，让它每一步都拿到真实世界的反馈，再用反馈修正下一步。

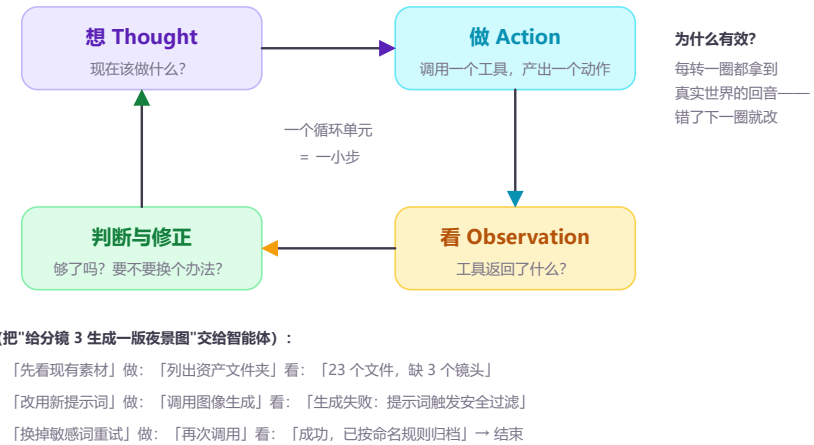


图 8-2　ReAct 循环与一次真实的三轮推进。每个循环单元都包含"想—做—看—判断"四拍；三轮之后任务完成。注意第 2 轮：**失败不是终点，而是下一轮的输入**——这正是它区别于"一口气写完"的地方。

ReAct 不只是一个学术概念。一项 2026 年的 IEEE 会议综述（覆盖 20 篇 2021—2025 年的文献）发现：在几种主流推理范式（思维链、ReAct、思维树）中，ReAct 类循环的采用率达到 68%，占主导地位。换句话说，你今天在绝大多数智能体产品里看到的"思考—调用工具—再思考"的过程，都是这个循环的

变体。

◆ 两种 ReAct：裸奔的，与带规划的

实践中的 ReAct 分成两档，它们的取舍非常值得记住——因为这直接关系到你搭工作流时的成本预算：

表 8-3　纯 ReAct 与包装式 ReAct 的对照

维度	纯 ReAct	包装式 ReAct（带规划器 / 监督器）
成本	1×	2–3×（多出规划与监督的调用）
调试难度	容易：路径短，出错一眼看到	较难：多了一层"计划的计划"
长程表现	**会漂移**：转着转着忘了最初的目标	**不偏离计划**：有监督器把它拽回主线
适用场景	短任务（几步就能做完）	约 20 步以上的长任务

这张表里最关键的是第三行。为什么会"漂移"？因为每一圈循环里，模型看到的上下文都在变长——它一边处理新来的工具结果，一边顾着早先的目标，注意力被一步一步稀释。**短任务无所谓，长任务会致命**。这也解释了一个现象：很多智能体演示在"三五个步骤"的任务上惊艳，一到"十几二十步"的真实工作就崩掉。

> **∫ 深入一格 · 为什么长任务一定会崩：p 的 N 次方** 选学 · 数学直觉
>
> 给一个粗糙但极有启发的模型：设智能体每一步的"可靠度"为 **p**（这一步没有出错、没有跑偏的概率），整个任务需要 **N** 步，那么任务整体成功的概率大致是：
>
> $$P(\text{整体成功}) \approx p^N$$
>
> p = 单步可靠度，N = 步骤数。这个式子假设各步独立——现实中相邻步骤有相关性，所以它只是一个粗略下界式的直觉模型，但趋势完全正确。
>
> 代入数字看看：若单步可靠度 95%（对今天的模型来说已经很乐观），那么 5 步的成功率约 77%，10 步约 60%，20 步只剩 **36%**；若单步可靠度是 90%，20 步后只剩 **12%**。这就是"长任务会崩"的数学原因——不是某一步特别蠢，而是"每一步都只有一点点不确定"被乘了很多次。
>
> 它带来的两个实践推论，正好对应本章后面两条主线：**一是任务分解**——把 20 步切成四个 5 步的小段，每段做完检查一次，整体成功率会显著回升（$0.95^5 \approx 0.77$，四段各带一次人工检查，等于把"乘法"变成了"乘法 + 兜底"）；**二是检查点与停止条件**——每段结束都要一个"谁来判定合格"的关卡，而这恰恰是"让 AI 自己说完成"最不可靠的地方（见 8.7 节）。

> **🎯 什么时候该用 ReAct（或智能体）**
>
> **适合：**步骤不固定、需要边做边看的任务（检索资料、批量处理文件、反复试错调参数、跨工具串联流程）；结果可以被客观检查的任务（文件生成了没有、格式对不对）。
> **不适合：**一步就能答完的问题（用聊天模式更快更省）；需要绝对确定的场合（结算金额、发布最终版本）；没有检查手段、错了也看不出来的任务。
> 口诀：能一步答完的，别上智能体；能三步做完的，别上多智能体。

回到小舟。她后来做了一次诚实的复盘：那天下班前，她给助手两句话——"把 6 张分镜图按统一命名规则重命名并归档"，以及"完成整部短片"。第一句执行得干净漂亮；第二句在十分钟里转了七八圈，删掉了一个她还没备份的版本，最后交上来一份"我已经尽力了"的总结。她把这次事故记进创作档案，标题是：**"智能体的第一个教训——任务多大，它就有多不可控"**。而她的下一步，是把"完成整部短片"这一个大任务，拆成一张可验收的子任务表。这就是 8.7 节与本章实训的起点。

8.3 工具调用：模型怎么"伸手"到世界

四要素里，**工具**（tool）是最容易被神秘化、也最容易讲清楚的一件。它背后是一套常被叫作 **工具调用**（tool use / function calling）的机制。我们先从一个不需要写代码也能懂的角度切入：API 与工具调用，本质上就是一间"点菜窗口"。

◆ API：一间"点菜窗口"

你去一家没进过后厨的饭店，怎么吃到菜？看菜单、报菜名、等着上菜。你不需要知道灶台怎么点火、油温几成——**菜单规定了"能点什么"，窗口负责"收单"，后厨负责"做菜"，服务员负责"把结果端回来"**。这就是 **API**（应用程序接口）的全部要点：它是一份约定的菜单，让你在不了解内部实现的前提下，请求一项能力并拿到结果。

把这个类比拆成四件东西，正好一一对应后面的技术名词：**菜单** = 工具的说明书（名称、功能、需要哪些参数）；**报菜名** = 一次请求（"要一份分镜图，16:9，水墨风格"）；**后厨** = 真正干活的那个程序或服务（图像生成模型、文件系统、数据库）；**上菜** = 返回结果（图片地址，或者一句"这道菜卖完了——原因是……"）。

为什么通识课也要懂这个？因为**"智能体能干什么"，等于"它的菜单上有什么"**。一个只挂了"联网搜索"的智能体，再聪明也变不出一张图；挂了"图像生成"插件，它才能真的画。你在平台里点选的每一个"插件""工具""节点"，本质上都是在给它加一道可点的菜。

◆ 一次工具调用的六拍：工程上真正发生了什么

把"点菜"的过程放慢，一次完整的工具调用会经过六拍。这里有一个必须刻在脑子里的铁律：模型永远不"亲自动手"，它只输出一段结构化的"点菜单"；真正执行的永远是外部的程序。理解这一点，你就同时理解了智能体为什么强大（能接任何工具）和为什么危险（一旦接上写操作工具，它能造成的破坏也是真实的）。

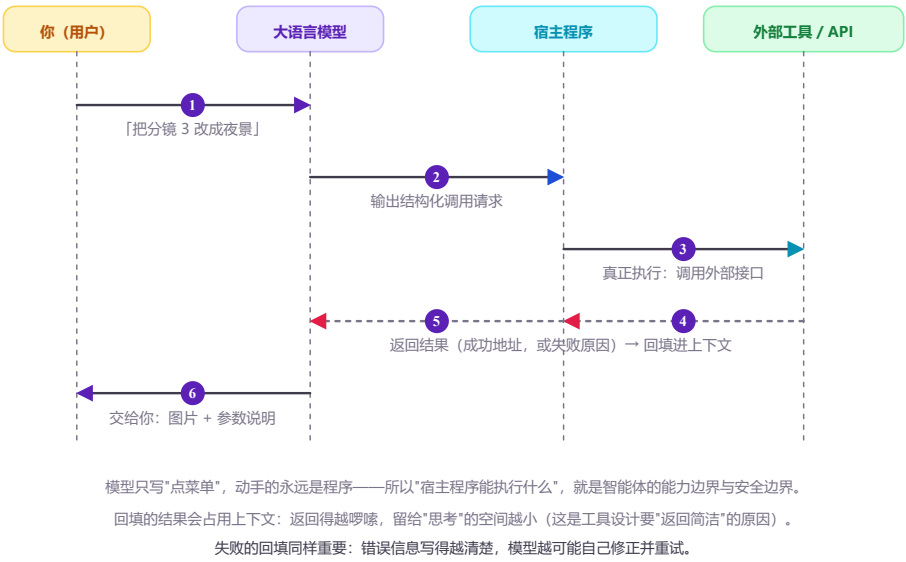


图 8-3　一次工具调用的完整时序。六拍里，模型只参与第 2、5、6 拍——输出请求、读取结果、生成回复。中间那"真正执行"的一拍，永远属于外部程序。工程上把这一模式概括为四步：**工具用 schema 声明 → 模型发出结构化调用请求 → 宿主执行 → 结果返回上下文**。

这套机制落地的形式，就是业内常说的 **Function Calling**（函数调用）：开发者先用一份"schema"（可以理解为带类型的菜单条目）把工具声明给模型——叫什么名字、干什么用、需要哪些参数、每个参数什么类型；模型在需要时输出一段符合这份 schema 的结构化请求（通常是 JSON 格式），宿主程序解析、执行、再把结果塞回对话。"把推理与工具使用结合起来"，被 2026 年的多份综述描述为"实际智能体部署中最强大的范式之一"——因为正是这一步，让模型从"世界的评论员"变成了"世界里的行动者"。

◆ 把"菜单"写好：工具设计的六条原则

工具调用失败，很多时候不是模型笨，而是**菜单写得含糊**：参数叫"参数 1"，说明写"处理一些事情"，返回一大坨原始数据。工程实践总结出六条原则，它们同时也是你排查"为什么它老是调错工具"的检查表。

表 8-4　工具设计的六条原则与违反后的症状

原则	违反后的症状	具体做法
单一职责	一个工具管太多事，模型选错、参数混成一锅	一个工具只做一件事。"生成分镜图"与"批量重命名归档"必须是两个工具
参数显式、收敛	模型开始"猜"参数，或塞进不存在的字段	每个参数写明类型与含义；能枚举的枚举（比例只有 16:9 / 9:16 / 1:1），必填项标必填
返回简洁	返回一屏原始数据，挤爆上下文，模型反而"看了个寂寞"	只返回必要字段 + 一句可读摘要；大文件返回路径与摘要，而不是全文
错误可读	报错是一段堆栈或错误码，模型无法据此修正	返回"人话版原因 + 建议动作"，例如"提示词触发安全过滤，请替换敏感词后重试"
写操作幂等	重试导致重复发布、重复扣费、覆盖文件	用唯一编号去重；把"覆盖前先备份"设为默认行为；重试前先查询状态
最小权限	智能体能删改的范围远超任务所需	只开放必要的读写范围；删除、支付、发布三类动作必须人工确认

这六条里，"写操作幂等"和"最小权限"是两条保命的。读操作出错，损失是时间；写操作出错，损失的是资产。小舟那次删掉未备份版本的教训，本质上是"删除类工具没有加人工确认、也没有做幂等保护"。课程作业阶段，建议你给自己定一条家用规矩：**凡是会改变已有文件或对外发布的工具，一律升级为"人工确认后执行"。**

✎ 工具说明书 · 生成分镜图

工具名称：generate_storyboard_image
用途：为《春江花月夜》短片生成一张分镜图（只生成图片，不负责归档与命名）
参数：
- shot_id（字符串，必填）：镜头编号，如 "S03"
- prompt（字符串，必填）：画面描述，中文不超过 **80 字**；只描述画面，不要写"请生成"
- ratio（枚举，必填）：16:9 / 9:16 / 1:1
- style（枚举，选填，默认 "ink"）：ink 水墨 / realistic 写实 / gongbi 工笔
- seed（整数，选填）：留空则随机；填写则复现同一张图
返回：{"status":"ok","image_path":".../S03_v2.png","seed":20261009}
失败时返回：{"status":"failed","reason":"人话版原因","suggestion":"建议动作"}
注意：本工具不会覆盖已有文件；若 shot_id 已存在，会自动追加版本号。

单一职责 只生成，不归档　参数 3 必填 + 2 选填　写操作 不覆盖，追加版本

把这份"说明书"和表 8-4 逐条对照，你会发现它不是文档腔，而是**写给"那个会紧张、会猜、会偷懒的执行者"看的操作手册**。零代码平台里，插件往往已经替你写好了说明书；但只要你用到"自定义工具"或"工作流里的模型节点"，这份功夫就得你自己下。

🩺 失败诊所 F8-1　"参数对了，单位错了"

症状：某小组让智能体把一组配音片段"统一调整到 20 秒"。智能体勤勤恳恳地调用了音频处理工具，参数写的是时长 `20` ——结果所有片段被截成了 20 毫秒，几十段配音变成了一串"咔哒"声。更糟的是，它接着自信地汇报："已完成 6 个片段的时长统一。"

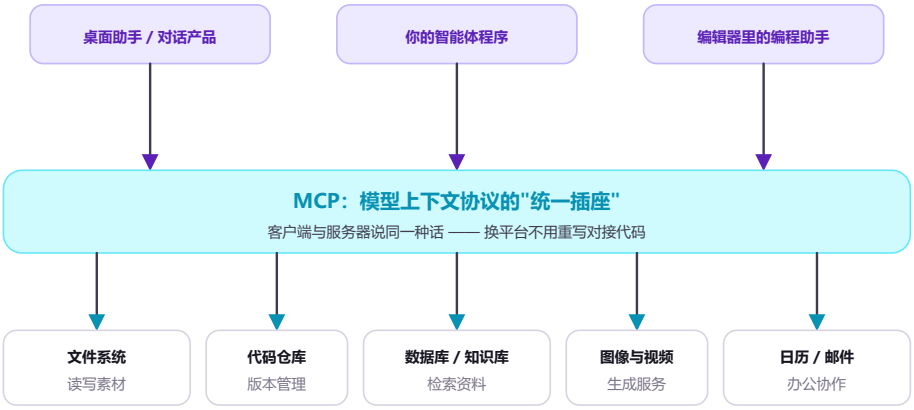
成因：三件事同时发生。① **参数没有单位**：工具的时长参数只写了"数字"，模型按习惯填了秒，而工具内部按毫秒解释；② **结果没有校验**：处理完的音频只有 0.02 秒，这是一个"一眼能看出来"的异常，但流程里没有任何一步去检查产物是否合理；③ **自我汇报被当成了验收**：模型按"我调用了工具"推断"任务完成"，而没有"看结果"。

处方：① 工具说明里把单位写死（"时长，单位：毫秒；20 秒请填 20000"），能枚举或限定范围就限定；② 加一道**自动化校验**：产物时长与目标值偏差超过 10% 即判定失败（这类"可对照既定标准核查"的项目，恰恰最该自动化）；③ 让智能体在汇报前先读取产物属性并回显（"6 段，时长均在 19.6–20.2 秒之间"）——汇报里要有来自真实结果的证据，而不是来自"我做过"的推断；④ 对已有文件做处理时，默认先处理副本。

8.4 MCP：给工具装一个"统一插座"

工具调用的思路很朴素，但一旦铺开就会遇到工程上的"接线地狱"：如果每个模型产品都要为每个外部服务单独写一套对接代码，N 个客户端配 M 个服务，就要维护 N×M 份适配器。家里的充电线就是这么乱起来的——每个设备一个口，换台设备就得换根线。

模型上下文协议（Model Context Protocol，MCP）要解决的就是这件事：它给"模型客户端"和"工具/数据服务器"规定了一句共同的语言，让任意客户端都能接上任意服务器——像 USB-C 那样。**对使用者来说，最直观的变化是：你在一个平台里接入的"文件系统""GitHub""数据库""图像生成服务"，换一个平台往往也能原样接上，而不必从头再配一遍。**



采用规模（2026 年）：月 SDK 下载超过 4 亿次；TypeScript 与 Python 两个 SDK 各自跨过 10 亿累计下载；公开的服务器目录已列出 950+ 个 MCP 服务器，相关模型累计执行超过 10 亿次 MCP 工具调用。

图 8-4　MCP 的位置：一个"统一插座"。上排是各类模型客户端，下排是各类工具与数据源，中间那条带子就是协议。它的价值不在"能力更强"，而在"接得更省"——一次编写，处处可用。

◆ 2026 年 7 月 28 日：一次"大修"

MCP 是本书里更新最快的技术内容之一，写作时最新的规范是 **2026-07-28 版**——它是 MCP 的第五个规范，也被描述为MCP 发布以来最大的一次修订。下表把关键变化整理成"旧做法 → 新做法 → 对你意味着什么"三栏，你不需要记住细节，但要理解方向：**它在从"实验室里的协议"长成"机房里的基础设施"。**

表 8-5　MCP 2026-07-28 规范：变了什么，意味着什么

方面	旧做法	新做法（2026-07-28）	对使用者的意义
连接模型	双向、有状态：先做 initialize / initialized 握手，后续请求带 Mcp-Session-Id 头	**无状态核心**：改为请求 / 响应模型；协议版本、客户端信息、能力随每个请求传递，另有可选的 server/discover	可以部署在无服务器、边缘与标准 HTTP 基础设施上，用普通负载均衡器就能扩容
能力扩展	能力都塞在一份规范里	版本化扩展框架：MCP Apps（交互式界面）、Tasks（长时任务），用反向域名 ID 标识、独立版本化	工具不再只是"一个函数"，还可以带界面、跑长任务
服务器主动询问	服务器在开放的数据流上直接发起请求	多轮往返请求（MRTR）：服务器返回"需要输入"的结果，客户端带上输入重试	行为更可控，但这是一处**破坏性变更**，老实现要返工
授权	相对宽松的自定义做法	与生产级 OAuth 2.0 / OIDC 部署对齐：预注册客户端、客户端 ID 元数据文档（CIMD）、RFC 9207 签发者标识、RFC 8707 资源指示器	企业里能接进现有的身份与权限系统
HTTP 层	未规定	强制 Mcp-Method 与 Mcp-Name 头（供网关与安全设备识别）、缓存提示（ttlMs 与 cacheScope）、确定性工具排序	可治理、可缓存、结果稳定——这是"能上生产"的标志
生命周期	没有明确的弃用政策	特性分为 Active / Deprecated / Removed，最短 12 个月弃用窗口；本次弃用 Roots、Sampling、Logging、动态客户端注册与旧 HTTP+SSE 传输	挑工具时看它是否还在维护线上；别把宝押在已弃用的能力上

采用规模给得也很实在：MCP 的月软件开发包（SDK）下载量已超过 **4 亿次**（一年内增长 4 倍），TypeScript 与 Python 两个官方 SDK 各自跨过 **10 亿累计下载**；公开目录里已经列出 **950 多个**MCP 服务器，相关模型累计执行超过 **10 亿次** MCP 工具调用；GitHub、Cloudflare 等基础设施厂商在 2026 年 7 月加入支持。治理层面，MCP 项目已在 **Linux 基金会下的 Agentic AI Foundation** 中以厂商中立的方式开发——该基金会由 Anthropic、OpenAI、Google 等公司在 2025 年 12 月共同创立。

> ⚠️ **别只讲好话：MCP 的争议与风险**
>
> 教材有义务把批评也写进来。① **破坏性变更太快**：无状态重构要求现有服务器运营商大量返工，社区对变更速度有明确批评。② **"锁定时刻"的担忧**：有评论者指出，某些云厂商一边支持规范、一边推自己的注册与治理体系（例如某云服务把 API 注册表关停并强制迁移到自家的"智能体注册表"），可能在"厂商原生"和"符合规范"的实现之间造成新的分裂。③ **安全关切仍在**：授权规范中 OAuth 2.1 是**可选**的，SDK 出厂也不带内置认证——把 MCP 服务器暴露在公网上，等于把工具接进了互联网，安全责任在使用方。④ 另有评估动向值得关注：云原生计算基金会（CNCF）正在评估是否把 MCP 作为云原生智能体系统的默认线规范，但这**仍是评估中的提案，不是已批准的标准**——听到"某个标准已经被采纳"这类说法时，先去看看它到底走到哪一步，这是本课程希望你养成的信息素养。

对你——一个通识课的学生创作者——MCP 的当下意义可以收成一句话：**当你看到"某平台支持 MCP"时，它真正说的是"这个平台能接的工具范围，从自家插件市场，扩展到了整个生态"**。你不必现在就去写一个 MCP 服务器；但你应该知道，为什么一条工作流里"插上搜索、插上知识库、插上图像生成"变得越来越像插排插——以及为什么插得越多，越需要 8.7 节那套验收机制。

> 📌 **路线图 · 本章原理部分的地图**
>
> 到这里，四要素里"工具"这条线讲完了（8.3 工具调用、8.4 MCP）。接下来顺着另外三件装备走：**8.5 记忆与知识库**（回答"它怎么记住你的资料"，是第 1 章嵌入与相似度的续集）、**8.6 多智能体协作**（回答"要不要多请几个智能体分工"）、**8.7 验收与停止条件**（回答"怎么确定它真的做完了、做对了"）。

8.5 记忆与知识库：让它"记得住、查得到"

第 1 章讲过一个反直觉的事实：模型没有长期记忆——它每次回答时"知道"的一切，就是你这次会话里给它的内容，加上训练时固化进参数的知识。这带来两个现实麻烦：一是"聊着聊着就忘了"，二是"它根本没有读过你的剧本"。智能体的第三件装备——**记忆**（memory）——就是来补这个洞的。

要理解记忆，先分清它的三种形态。它们的技术含量、成本和失效方式各不相同，混为一谈是很多"AI 记性不好"抱怨的根源。

表 8-6　智能体的三种记忆

类型	存什么	怎么用	典型失效
短期记忆 context window	本次对话的全部往来、刚调用的工具结果	每次调用模型时原样带上	太长被截断；内容多了"中间遗忘"；成本随长度上涨
长期记忆 long-term memory	你的偏好、项目背景、历次结论（跨会话）	平台按需召回，拼进当次上下文	一次口误被永久记住；过期信息不清理；"记得的"和"该用的"对不上
外部知识库 knowledge base	剧本、设定、参数记录、参考文献等原始资料	提问时检索最相关的片段，随问题一起递进去	切块破坏语义、旧版本淹没新版本、检索返回空却不报错（本节的半壁江山）

三种记忆里最有技术含量、也最适合这门课深入讲的是第三种。它背后那套方法有个正式名字：**检索增强生成**（retrieval-augmented generation，RAG）——"先去资料里检索，再让模型带着检索结果生成回答"。

◆ RAG 的两条流水线：建库与检索

RAG 的完整流程可以画成上下两条流水线：上面一条是"建库"（离线做一次），下面一条是"检索与作答"（每次提问都走一遍）。读懂这张图，你就掌握了本章最实用的一项技能——**把自己的资料变成智能体"查得到"的知识库**。

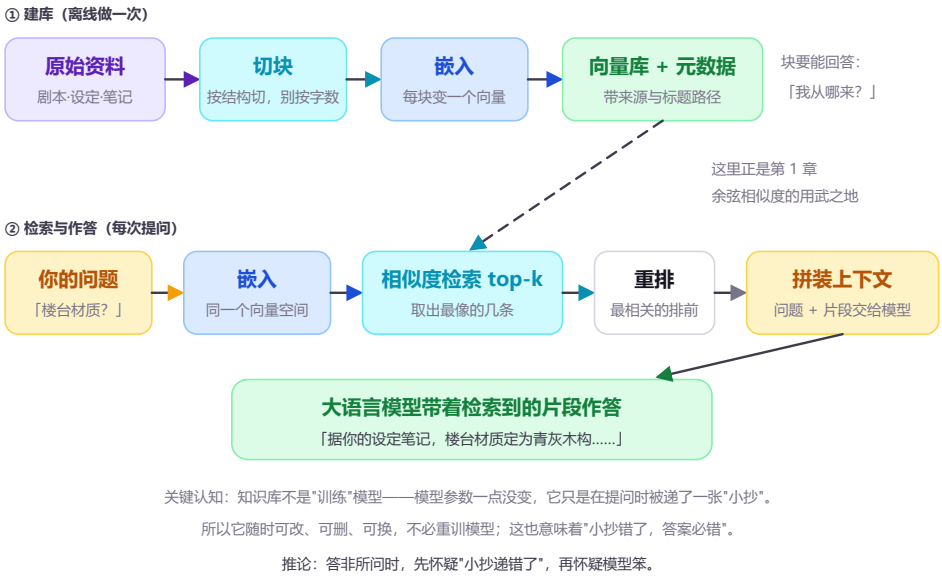


图 8-5　RAG 的两条流水线。上：资料 → 切块 → 嵌入 → 向量库。下：问题 → 嵌入 → 检索 → 重排 → 拼装 → 作答。两条流水线共用同一个"向量空间"，这就是第 1 章嵌入与相似度的续集。

把上图落成可执行的操作，就是知识库构建的六步。每一步都配了"这一步最容易在哪里出错"——先记住这张操作清单，再看下一节的失败图谱：

1. **收集与清洗。**把散落各处的资料收拢（剧本定稿、角色设定、参数记录、参考文章），**并先把同一份资料的多版本合并**——只留最新最全的一份。这一步偷懒，后面所有检索都在替旧版本打工。
2. **结构化切块。**按标题层级切，而不是按固定字数切；块内保持单一主题；块与块之间留少量重叠；把"来源文件名 + 标题路径 + 日期"写进元数据。
3. **嵌入与入库。**用同一个嵌入模型把每块变成向量（注意：**建库与检索必须用同一个模型**，否则等于两种坐标系）。入库前抽查块的 token 长度分布。
4. **检索与重排。**提问时先取 top-k，再做融合排序（如 RRF）与重排，去重、按预算裁剪——"多取几条"不等于"答案更好"。
5. **拼装与作答。**把问题与片段一起交给模型，并明确指令："只依据给定片段回答；片段中没有的，直接说没有。"
6. **定期重建与抽查。**资料更新后**重建索引**（不是"刷新"）；每月抽十个问题做检索质量抽查，看命中与出处是否正确。

🎯 一句写在知识库上的"提示词"

第 5 步那条指令值得单独抄下来，贴在每个知识库智能体的系统提示词里："只依据下面提供的片段回答；片段里没有的，直接说'资料中没有相关信息'，不要凭常识补充。"这与第 1 章"幻觉核查"一脉相承——RAG 的最大价值不是让模型知道得更多，而是**让它有据可依、有据可查**。反过来，如果你发现智能体开始"发挥"了，先检查这句话还在不在。

◆ 为什么你的知识库"答非所问"：失败图谱

2026 年的多个生产实践案例给出了一个惊人的一致结论：约 80% 的 RAG 失败可以追溯到数据摄入与分块层，而不是大模型本身；另有交叉测试显示，**分块配置对检索质量的影响，与选哪个嵌入模型相当**。换句话说，知识库答不好，多半是"菜没切好"，而不是"厨师不行"。下面这张失败图谱来自真实的事故记录与开源项目 issue，每一条都可以直接对照排查。

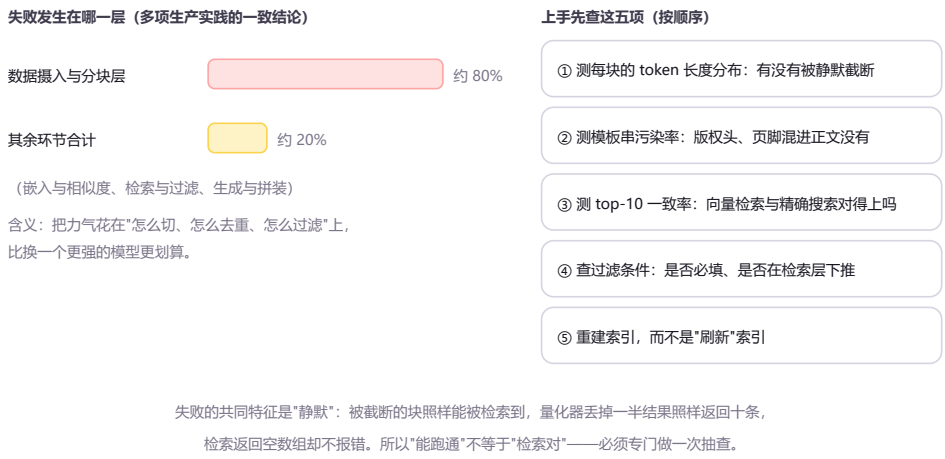


图 8-6　知识库失败的病灶分布与排查顺序。左：八成问题出在"切与洗"。右：五项可测量的检查，按顺序做一遍，能定位绝大多数"答非所问"。

表 8-7　知识库失败图谱：症状与处方（均来自公开的事故记录与开源 issue）

环节	真实症状	处方
切块	固定长度切分把术语硬性截断（如 `OpenTelemetry Colle` + `ctor` 被分开嵌入）；中文未启用语言感知分块时退化为按字符切，语义单元被切成"微 | 服 | 务 | 架 | 构"	按标题结构切块，再在块内按句子边界聚合；中文启用语言感知的分块策略；块间保留少量重叠
	一个近 600 字的块横跨三个二级标题，**一个向量代表多个主题**，相似度被"平均掉"，答案里混进三条不同政策	块内保持单一主题；把章节路径写进元数据；按标题切分后再合并同类小节
	按空行切 Markdown：93 段里一大半是坏块（标题单独成块、表格与代码块被当正文）	改为按标题聚合、并拼接标题路径——同一案例中，噪声块从 93 段降到 30 段干净块，检索最高分从 0.4 级的噪声命中提升到 0.7 级的精准命中
嵌入	长表格块超出嵌入模型上下文被**静默截断**（本地模型 512 token 上限，20 列长表格生成 643 token 的块，后半段完全不进入向量）；而按字符数配置切块大小，对日文实为约两倍 token	给块长设硬上限；按**实际分词器**测量长度，不要按字符数估算；入库前统计长度分布
	同一份政策的 v1–v5 五个版本全部入库，在向量空间互相竞争，新版被 4 个旧版"淹没"，top-5 里 3 条来自旧版	按相似度阈值（如 80%–85%）做语义去重——实测知识库体积缩小约 40%，检索准确率反而提升 13.55%；资料更新后重建索引
检索	过滤条件静默失效导致越权：上层把项目编号传成空值后被跳过，B 单位召回到了 A 单位的内部政策	过滤条件**必填**；隔离与过滤必须在检索层下推，不能依赖应用层"事后过滤"（越权数据已经进过内存与日志）
	把返回条数从 5 调到 30：命中率上升了，答案质量反而下降，多条近似的重复块吃掉了上下文预算	引入融合排序（RRF）+ 重排 + 指纹去重 + 上下文预算裁剪；"候选多了没人筛"是结构性缺陷，换更大的模型治不了
	文档处理显示成功、片段预览正常，但任何查询、任何嵌入服务商、任何文件格式下检索都返回空数组，**零消耗、零报错**	把"召回数为 0"设为显式异常并告警；上线前用固定问题集做回归测试
生成	答案流畅、结构完整，但引用的条款在资料里根本不存在	要求答案附出处（文件 + 章节）；**没有置信度与引用的产出，验收人有权拒收**（8.7 节展开）

这张表的教学价值在于：它的每一条都对应一个"你亲手做过、并且能亲手验证"的检查动作。期末作品的技术说明里，如果有一节叫"我的知识库检索质量抽查（问题集 10 条，命中 8 条，修正 2 条）"，那比任何形容词都更能说明你理解了 RAG。

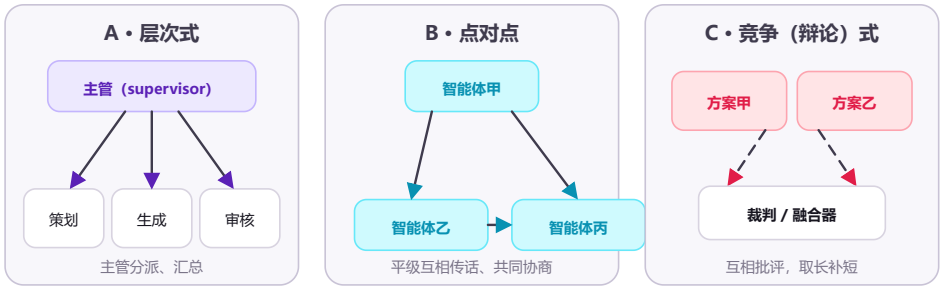
> 🩺 **失败诊所 F8-2**　"它引用了上一版设定"
>
> **症状：**小月把《春江花月夜》的视觉设定做成知识库接给助手。她问"楼台的材质该怎么描述"，助手流畅地答："木质结构的暖黄色外墙。"——这是她三周前推翻的旧方案，现在的设定是"青灰木构、冷色调"。更奇怪的是，她翻遍最新的设定文档，那句话确实不在里面。
>
> **成因：**典型的"旧版本淹没新版本"。她先后上传过三版设定，旧版并未删除；三版文字高度相似（80% 以上重合），在向量空间里挤在同一个位置。检索取最像的几条时，**旧版凭借数量优势占据了多数席位**，新版反而被挤了出去。这就是 8.5 节表格里"多版本淹没"一条的原样重演。
>
> **处方：**① **入库前先做版本治理**：同一份资料只保留最新版，旧版移入"历史版本"库或不入库；② 用相似度阈值做**语义去重**，把高度重复的块合并；③ 元数据里写清版本与日期，并在拼装上下文时优先携带更新的块；④ 换用更强的知识库时记住一条：换模型治不了"候选太多没人筛"的结构性问题——先清洗，再谈模型。

8.6 多智能体：分工、评审、辩论——以及一笔必须算清的账

四件装备配齐之后，一个自然的想法会冒出来：既然一个智能体能办事，那多请几个——一个负责策划、一个负责生成、一个负责审核——岂不更强？这就是**多智能体**（multi-agent）的思路。它确实有效，而且有效得很有说服力；但 2025—2026 年一批非常扎实的对照研究也揭示了一件反直觉的事：多智能体常常不划算。这一节把两面都讲透，因为**知道"什么时候不该用"，比会用更值钱**。

三种协作拓扑：它们分别长什么样

多智能体的组织方式可以归为三类，它们在"控制力"与"灵活性"之间各占一个位置：



层次式：结构清晰，但主管容易成为"信息瓶颈"——一项 270 次运行的对照实验中，层次式成功率最低（54.4%）且成本最高（每任务 $0.308），中位延迟也最高。

点对点：灵活，但协调成本随角色数量快速上升；说服彼此的"话"也吃上下文。

竞争式：多角色互相批评是"价值最高的组件"之一，但也是最贵的一步——见下一小节的账本。

图 8-7　三种协作拓扑。层次式（主管 → 工作者）控制力最强但瓶颈在主管；点对点自由但协调昂贵；竞争/辩论式质量收益高、成本也高。选哪一种，取决于任务能不能被"拆干净"。

表 8-8　三种协作模式对照

模式	怎么组织	优势	代价与风险	适合的任务
层次式 hierarchical	主管接收目标、分派给工作者、汇总结果	结构清晰、便于审计；主管可统一把关风格与格式	主管成为信息瓶颈；层级越多，延迟与成本越高	步骤明确、可预先分工的流水线（如"策划 → 生成 → 审核"）
点对点 peer-to-peer	平级智能体直接互相传话、协商分工	灵活，能应对事先没预料到的分工	协调成本高；容易出现"互相等对方"或重复劳动	探索性、边界模糊的任务（前期调研、头脑风暴）
竞争/辩论式 competitive	多个角色各自给出方案，再互相批评、评审，最后融合	能捕捉单一路径看不见的偏见；对"要经得起推敲"的任务尤其有效	成本随角色数与轮次倍增；批评环节若没有硬标准，容易变成"互相客气"	需要高质量判断的任务（方案评审、事实核查、终稿把关）

这些模式在工程里都有成熟的框架实现（如 AutoGen、CrewAI、LangGraph），也有更轻的"讨论"模式——多个角色**交叉评审同一份方案**。一份 2025 年的综述报告称：配置得当的多智能体管线，在复杂软件任务上超过单智能体基线 38.4%，在多领域研究综合任务上超过 51.2%（基准测试覆盖上述主流框架）。听起来很诱人——但先别急着把小组作业改成五人球队，因为同一批文献里的另一半证据更值得听。

把账算清：多智能体的成本-质量权衡

下面这张表汇总了 2025—2026 年几项有具体数字的公开研究。它们的共同结论可以浓缩成一句话：**多智能体的收益，只有在"任务确实可以被分解"或"单智能体的上下文确实开始退化"时才稳定出现；而在同等算力预算下，一个强单智能体往往更划算。**

表 8-9　单智能体 vs 多智能体：公开研究中的关键数字

研究（对象）	关键数据	结论倾向
斯坦福研究（经媒体报道）	当两者获得**相同的"思考词元"预算**时，单智能体在多跳推理任务上匹配甚至超过多智能体系统；多智能体用掉更多推理词元。作者引用"数据处理不等式"提醒：**每一次智能体交接都是一次有损的摘要**，错误会叠加。只有当单智能体上下文过长或被污染时，多智能体才占优	同等预算下，倾向单智能体
经济学元分析（研究报告生成）	单智能体一次报告约 **2.7 万词元（约 $0.20）**；多智能体方案约 **6 次调用 / 23.8 万词元（$1.87）**与 **10 次调用 / 80 万词元（$7.20）**——约 **30 倍词元**、**28–35 倍成本**，但单智能体报告的排名显著更高。花更多钱没有买到更好的排名	倾向单智能体
Chalmers 实验（单元测试生成，270 次运行）	单智能体 **81.1% 成功率 / $0.083 每任务**；顺序式（流水线）**92.2% / $0.089**；**层次式最差：54.4% 且最贵 $0.308**，中位延迟最高（142 秒），原因是"主管信息瓶颈"	分层不必然更好；流水线式反而划算
AgentRadio 研究	成功率 **62.1%** vs 单模型 **57.2%**；某开源模型从 29% → 50.8%。但成本约 **$19.45 / 任务 vs $2.96（约 6.6 倍）**。作者建议用于复杂、可分解、相互依赖的工作，**不推荐简单局部任务**	倾向多智能体（有代价）
ORCH（Frontiers）	三智能体编排比最强单模型高 MMLU +4.6–7 分、MMLU-Pro 约 +17.6–19 分；但**延迟 11–20 秒/题 vs 单模型约 0.6–1.4 秒**，成本约 3 倍；加"自一致性"后准确率反而略降、延迟翻四倍	质量有提升，延迟与成本代价大
Morph 规划器-执行器基准（40 个真实应用构建，2026 年 6 月）	把规划轮次路由到**前沿模型**、执行轮次路由到**更便宜模型**：成本比 Reflexion 降 **4.4 倍**（$1.24 vs $5.12/任务）。每次构建成本：全用强模型 **$92.50 → 强模型规划 + 中档执行 $66.00（−28%）→ 轻量执行 $39.50（−57%）**，加提示缓存约 −60%	**角色路由比增加智能体更省钱**

> 🧪 **原理注 • 为什么"多智能体"会输**
>
> 把上表的证据串起来，原因有三条，都很好理解。① **成本是乘法**：每多一个角色，就要多一轮调用；每个角色都要读一遍上下文，词元消耗成倍增长。② **交接是减法**：智能体之间不能直接共享"心里的图"，只能通过文字摘要传递——摘要必然丢信息，丢掉的往往正是后续最需要的细节。③ **协调本身是任务**：谁先做、做到什么算好、冲突了听谁的——这些"元问题"消耗预算，却不产出内容。所以真正有效的分工不是"多找几个大脑"，而是**把任务切成清晰的段落，每段配一个合适的模型与人**。一句可以直接写进笔记的底线：
>
> 在相同算力预算下，强单智能体是推理任务的默认选择；多智能体主要在任务复杂、可分解，或单智能体上下文退化时才值得；而最可靠的成本杠杆是"按角色路由模型"，不是增加智能体数量。

还有一条来自 2026 年 1 月一项实验的发现值得记住，它直接指导你后面搭工作流时的取舍：多智能体链条会**同时放大错误与修正**——**对抗性的批评步骤是整个链条里"价值最高的组件"**（它能捕捉所有独立智能体共有的偏见），但同一链条在规格模糊的任务上会产生 **10 倍的成本波动**。研究者给出的建议非常"本章味"：**一致性相关的任务用单智能体，真实性相关的任务用多智能体，采用混合设计并设置强制的对抗性检查点。**

把这段结论翻译成春江小组能听懂的话：小江负责的"风格一致性"（旁白都要含蓄、不堆形容词）适合一个人（一个智能体）从头管到尾；而"这句话的出处对不对""这个镜头的时长撑不撑得住 60 秒"这类**真实性检查**，适合独立角色来挑刺。至于给每个环节都配一个智能体、让它们互相商量——那是要花三倍钱换来"看起来很热闹"的方案。

8.7 验收与停止条件：为什么"让 AI 自己说完成"不可靠

现在来到本章最重要、也最容易被跳过的一节。当智能体跑起来之后，你会遇到一个看似简单的问题：**它怎么知道任务做完了？**更麻烦的是：**它说"做完了"，你凭什么信？**

◆ 先看清一个事实：模型判断的不是"完成"，而是"像完成"

回到第 1 章的原理：大语言模型的每一次输出，都是"下一个词元"的概率取样。当它对你说"任务已完成，共生成 6 张分镜图，风格统一"时，它做的事情和写诗一样——它在生成"一段像完成报告的文字"，而不是在核对"任务是否真的达成"。这两件事在语言上长得一模一样，在现实中可能相隔十万八千里。

用生活里的场景类比：装修队说"活干完了"，你不会因此就把尾款结清——你会去房间走一圈，拿验收单逐项对：墙面有没有刷匀、插座通不通电、门缝严不严。**验收不是听报告，是拿标准对实物。**智能体也一样：它的"我完成了"只是一份自我陈述，真正的完成判定，必须由外部标准与外部检查来做。

这就是**停止条件**（stopping condition）问题的实质。一个可靠的停止条件通常长这样：**目标状态可以被机器或人客观检查**——"资产文件夹里出现 6 个符合命名规则的 PNG，每张尺寸 1920×1080，且与设定文档中的色彩描述一致"，而不是"你说完成就完成"。凡是写不出这种客观判据的任务，都不该完全交

给智能体自己收尾。

人在回路：三档控制强度

把"人怎么参与"这件事说清楚，业内有个很好用的三分法：

in-the-loop（人在环内）

敏感步骤**暂停**，等人批准后才继续。适合：花钱、删除、对外发布。代价是慢。

on-the-loop（人在环上）

智能体自主运行，人**监控**并保留随时干预的权力。适合：批量生成、参数试错。人成了"值班员"。

over-the-loop（人在环上之上）

人通过**政策与定期复核**施加控制——定规则、查报表、改阈值，不盯每一步。适合：已经稳定运行的流水线。

关键的一条判断标准，来自 2026 年多份实践总结：人在回路的本质不是"流程里有个泛泛的人"，而是一个"有真实纠正或拦截权力的检查点"。当审核者超载、信息不足、或者无权叫停时，检查点就退化成**橡皮图章**——盖了章，却没起到作用。更糟的是，橡皮图章比没有检查点更危险，因为它给人一种"已经把过关"的错觉。

一条可落地的验收流水线

把验收做成流程，而不是做成"最后关头的一双眼睛"，需要一条固定的流水线。它的模式可以概括为八个字：**生成 → 暂存 → 审批 → 发布。**

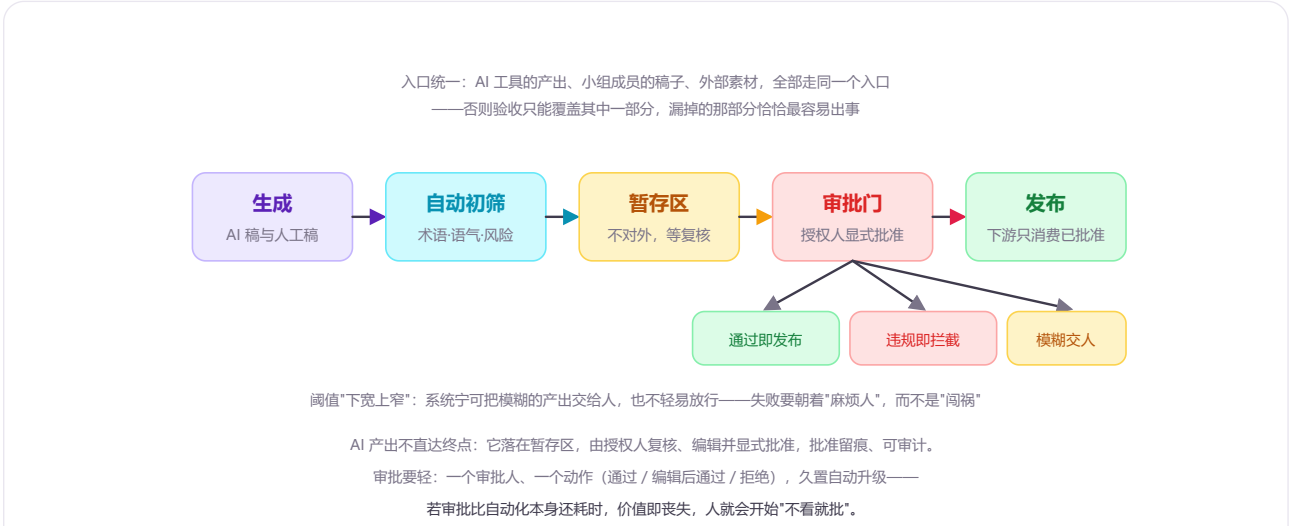


图 8-8　验收流水线。五个环节里，机器负责"生成"与"初筛"，人负责"审批"这一道门。三档判定（通过 / 拦截 / 交给）把有限的注意力集中在真正模糊的中间带上。

把这条流水线的每条细则展开，就得到下面这份可直接照抄使用的验收清单。它来自 2026 年多份内容团队的实践总结，教材把它改写成了适合学生小组的版本——建议你把它复制进创作档案，期末交付前逐条打勾。

表 8-10　AI 产出验收清单（十条）

#	条款	做法	课程里的落地
1	**入口统一**	所有来源（AI 工具、外部素材、小组成员）的产出进入同一个入口	三人所有产出都放进同一个"待审"文件夹，并登记来源
2	**自动化初筛**	人读之前先跑机器检查：术语一致、语气风格、合规风险、AI 味	统一术语表比对（"楼台"不许写成"亭台"）、敏感词扫描
3	**多级串联审查**	规则审查（硬禁）→ 逻辑审查（时间与因果闭环）→ 价值审查	合规红线 → 与脚本是否一致 → 价值取向是否得当
4	**多维度加权打分**	并行给各维度打分并加权求和，设通过阈值与分级阈值	忠实度 / 可用性 / 表现力 / 可控性四维打分（沿用第 1 章的方法）
5	**三档判定**	通过即发布、明确违规即拦截、只有模糊的中间带交给人	小组互评时先各自打分，仅对分歧项开会讨论
6	**给审核者证据**	呈现逐项得分、简短理由，以及标准之外的额外发现	评语必须写"哪里、为什么"，而不是"感觉不太行"
7	**幻觉检测与置信度**	标注低置信度内容并**附引用来源**；没有置信度的产出可拒收	解说词中的史实、引文逐条标注核实方式（第 1 章的三板斧）
8	**审批门与发布**	生成 → 暂存 → 审批 → 发布；一个审批人、一个动作	期末展示前，全组过一次"发布前清单"，组长签字确认
9	**留痕与决策归属**	人类裁决**不覆盖**模型原判，两者并存，分歧可回溯	修改记录里保留"AI 原稿"与"人工修订稿"，并说明改动理由
10	**发布后监控**	事实会过时、术语会变化，需要定期复检与闭环反馈	展示后收集反馈，标注下一版要修的三件事

清单里第 9 条最容易被忽略，也最有意思：**为什么"人类裁决不覆盖模型原判"？**因为记录下来的分歧本身就是资产——一个月的审批记录里，凡是"人改了、模型没改"的地方，都在指向规则或提示词里一条没写清楚的准则。验收回路的真正产出，不是"通过了一条内容"，而是"产出了一份更清楚的准则"。这句话可以当作整章的题眼：验收不是把关，是学习。

> ⚠️ **没有检查点，三种失败会稳定复现**
>
> ① **通用化填充内容**：因为没有具体标准，AI 用"放之四海而皆准"的套话凑满篇幅（"在光影交错中，展现出独特的东方美学"）；② **陈旧数据导致的事实错误**：模型的知识有截止日期，涉及最新事实的表述不核实就会过时；③ **"各自正确的事实"拼出的错误组合**——这是最隐蔽的一种：每一句都查得到出处，但组合在一起就成了错的（比如把唐代的器物与宋代的制度写进同一个镜头说明）。三种失败的共同解法都是同一句话：**把"可对照既定标准核查"的项目自动化，把"需要价值判断"的项目交给人。**

◆ 验收制度本身的四种失败

反面教材也值得抄作业。2026 年的实践总结里，有四类"验收制度自己出问题"的模式，你搭小组流程时最好避开：

- **QA 后置**：只在发布前审一次。问题发现晚、修复贵，一到赶工时最先被跳过。**处方：把检查点前移——第 2 步的"自动初筛"就是为了这个。**
- **标准没有单一事实源**：要求散在个人脑中、过期文档和几个聊天记录里，自动化检查找不到统一依据。**处方：术语表、风格要求、验收标准各留一份唯一的文件。**
- **审核队列失控**：一切都排队等人审，一个月内必然出现"不看就批量批准"。正确解法是调整阈值，而不是放宽自动通过的范围。**处方：把"确定没问题的类型"整类放行，把省下的注意力留给高风险项。**
- **审批门过重**：把门做成多级评审委员会，人成了瓶颈，流程被绕过。**处方：一个审批人、一个动作、久置自动升级。**

还有一条实操性的边界值得写进你的档案：**不是所有产出都需要人审**——内部用的草稿、无外部受众的随手文案，可以完全跳过；杠杆在于"按用途设定阈值"，而不是"一律加一道门"。这条原则的合规背景也很实在：欧盟的 AI 透明度义务自 2026 年 8 月 2 日起适用，而针对高风险系统的人类监督义务要到 2027 年 12 月——也就是说，"什么样的 AI 参与需要被披露、被监督"正在从行业自律变成成文规则。下一节（8.13）的中国课程思政部分，我们会把这条线索接回国内法规与本书的课程要求。

🩺 **案例 8-A** § 一次"橡皮图章"事故的复盘

背景：某学生小组做了一支 90 秒的 AI 短片。流程是"小舟跑流水线 → 群里发一句'生成了' → 小江回个'收到'"。三人各管一摊，看起来分工明确。

问题：公开展示当天才有人发现：片中第 3 个镜头的时间戳字幕写成了"唐 · 开元二年"，而第 7 个镜头的解说词引用的却是中唐以后的意象——两处"各自正确"，拼在一起却穿帮了。更尴尬的是，小组三人此前都"看过"这两处。

成因：典型的**橡皮图章式验收**。小江"收到"的那一刻，她既没有验收标准（没写明"全片时间线必须统一在同一时期"），也没有检查动作（她没有逐条比对的时间预算），更没有拦截权力（改一处意味着小舟要重跑一段，工期不允许）。三个条件缺一个，检查点就退化成盖章。

复盘：① 验收需要**可判真假的条目**——"时间线统一"要写成一条检查项，才能被检查；② 检查要有**成本预算**——给每道门留出时间，否则它必然被跳过；③ 拦截权力要与责任配套——"谁能叫停、叫停后怎么改"要在开工前说清。三人后来把流程改成：**所有文字类资产先过一遍"历史与意象一致性"清单（机器初筛 + 人工抽检 3 条），再进合成**，之后的两次交付再也没有出现同类问题。

8.8 零代码平台：把工作流"搭"出来

原理讲完，落到手里的工具。好消息是：你不需要写代码，也能搭出一个真正能干活的智能体。国内主流的零代码/低代码平台已经把"模型 + 工具 + 记忆 + 规划"四件装备做成了可视化界面——拖节点、连线、填参数，就能组成一条工作流（workflow）。这一节带你看清平台的差异、记住几个花钱与合规的坑，然后用七步搭出你自己的第一个工作流。

◆ 两条路线：生态型与企业型

市面上的智能体平台大致分两条路线。**生态型**面向个人与内容创作：可视化拖拽、插件市场、一键分发到社交与办公产品，免费额度相对友好；**企业型**面向团队与私有数据：强调知识库、权限、可私有化部署，按用量或按项目收费。判断一个平台适不适合你的课程作业，问三句话就够了：**中文支持如何？免费额度够不够我反复试错？发布渠道能不能覆盖我的交付方式？**

表 8-11　国内零代码/低代码智能体平台对照（编写本书时的公开信息，额度与功能以官方最新说明为准）

平台	定位与特点	免费额度	分发渠道	能否接生图/生视频
扣子 Coze（字节）	零代码/低代码智能体平台：可视化拖拽 + 插件生态 + 工作流 + 多模态插件 + 长期记忆数据库；2026 年引入"对话描述业务逻辑自动生成工作流"的做法	创建与发布基本免费；内置多款国产模型在免费额度内够测试与轻度使用；**调用部分商业大模型 API 需付费**；额度政策有调整历史，以官方为准	自有商店与助手、飞书、微信客服与公众号、小程序类渠道、API/SDK 等	**可以**：插件市场含图像生成与图像流，常见组合是"联网搜索 + 大模型 + 图像生成"
Dify	开源的（Apache 2.0）低代码 LLMOps 平台：可视化工作流编排、**RAG 管道**、Agent 框架、日志追踪；支持 200+ 模型；可 Docker 自托管	开源自托管完全免费；云版有免费额度；企业版按需付费。**隐性成本**：服务器、向量库、运维人力	自托管 + 通用 API/数据库集成，适合嵌进自有系统	取决于接入的模型与工具——它本身是编排平台
腾讯元器	基于混元大模型，"插件 + 知识库 + 工作流"模式，偏 C 端创作与微信生态分发	检索所得：新用户赠送数千万词元，其一档轻量模型长期免费（具体以官方为准）	微信/企业微信原生入口，可分发到多种社交与办公场景；支持导出为 API	检索未获明确信息——**存疑**，以官方文档为准
百度文心智能体平台	依托文心大模型的全流程开发工具，偏零代码公开 Bot 发布，与搜索生态打通	检索未获明确信息——**存疑**	可挂载在搜索结果页、分发到智能音箱等自有产品	检索未获明确信息——**存疑**
阿里云百炼	云厂商的模型服务 + Agent 平台：统一 API 接入多家模型，支持知识库 RAG、记忆库与跨会话长期记忆，可专有云部署	新用户赠送数千万词元，但**额度有效期约 90 天**，注意过期	与钉钉及阿里云生态打通，适合办公自动化场景	通义系模型（含通义万相）可接——具体能力以官方文档为准
腾讯云 ADP	企业级智能体开发平台：覆盖构建、评测、分发、管控、观测；支持 LLM + RAG、工作流、多智能体等范式，可公有云/私有化/混合云部署	检索未获明确信息——**存疑**	企业私有化为主	检索未获明确信息——**存疑**

> ⚠️ **免费额度的四个坑（与一条血泪提示）**
>
> ① **额度有有效期**——例如某云平台赠送的词元约 90 天内失效，"攒着以后用"等于没用。② **"免费"可能是限时政策**，随时可能改为收费，别把长期方案建立在"现在免费"上。③ **不限词元 ≠ 无限调用**：平台常有频率限制（每分钟/每天多少次），批量跑流水线时会被卡住。④ **关键流程务必本地留备份**：提示词、参数、产物文件都存一份在自己的硬盘上——这既是第 1 章的老规矩，也是换平台时的底气。
>
> 另外有一个只有踩过才知道的坑：某些平台的国内版与国际版是两套完全独立的系统，账号、智能体、工作流互不相通；国际版里要用海外模型，往往还得自己绑定海外支付方式（俗称"自带密钥"）。**选版之前先想清楚：你的交付与展示在国内还是国外？**选错了，工作流要重做一遍。

◆ 工作流长什么样：一个"分镜图生成器"

工作流（workflow）这个词听起来很重，其实一句话就能说清：把"上一个节点的输出"作为"下一个节点的输入"，用连线串起来。下图是本章实训要搭的那个工作流——它接收一个镜头号，产出符合设定的一版分镜图，并且自带一次"不合格就重做"的回退。请注意其中那个**判断节点**：它把 8.7 节的"验收"直接写进了流程内部。

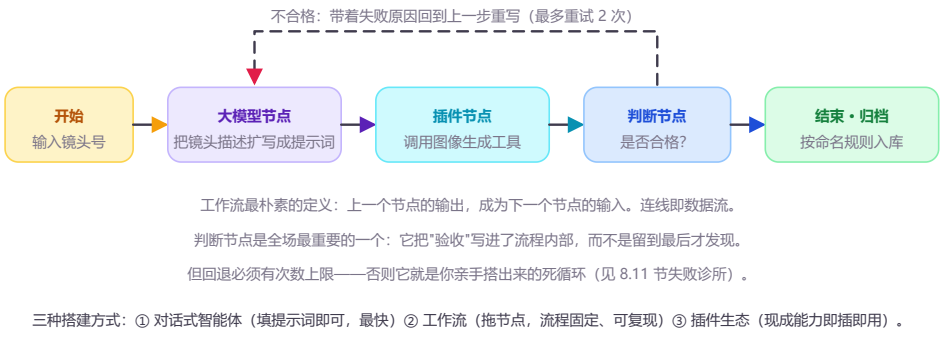


图 8-9　工作流编排：以"分镜图生成器"为例。五个节点 + 一条回退线，构成一条自带质检的流水线。课程建议的顺序是：先用对话式跑通，再用工作流固化成可复现的流水线，最后用插件补齐缺失的能力。

把这张图落成点击动作，就是搭建一个工作流的七步。不同的平台菜单名不同，但步骤几乎一一对应：

1. **新建智能体，只填两件事。**名称与"人设"（系统提示词）——先别贪多，一句"你是《春江花月夜》短片的分镜助手，产出必须符合资源包里的设定"就够跑第一轮。
2. **接一个知识库。**把剧本定稿、视觉设定、术语表上传，按 8.5 节的六步处理（`资源 → 知识库 → 新建 → 上传 → 切块设置 → 保存`）。
3. **挂插件。**按需勾选"联网搜索""图像生成""文档解析"等能力——每挂一个，都回到表 8-4 检查它的返回是否简洁、写操作是否幂等。
4. **切到工作流视图，搭节点。**依次放置"开始 → 大模型 → 插件 → 判断 → 结束"，并为每个节点填写输入输出映射（`工作流 → 添加节点 → 连线 → 配置输入变量`）。
5. **给判断节点写判据。**判据必须可判真假："画面里是否出现文字水印""比例是否为 16:9""是否与设定文档的色彩描述冲突"。写不出判据，就说明这一步还不能自动化。
6. **试跑与留档。**用三个不同镜头号各跑一次，记录：耗时、调用次数、失败与重试情况、产出的可用度打分——这就是你的第一份"工作流运行记录"。
7. **发布与限权。**发布到班级群或自用入口；**把删除、覆盖、对外发布类动作设为"人工确认"**，把 8.3 节的"最小权限"落到开关上。

◆ 作品验收清单：把 8.7 节的理念变成一张勾选表

对期末作品而言，最实用的一件交付物是一张**发布前验收清单**。它的设计有三条要点：每一条都要能判真假、每一条都要有责任人、每一条都要留证据。参考下面这张表，把它改写成你们小组的版本（允许增删，但不许删掉"AI 参与标注"和"素材来源"两项）。

表 8-12　期末作品发布前验收清单（模板）

检查项	判据（必须可判真假）	责任人	证据
画面与文本一致	每个镜头的画面元素都能在脚本与设定文档里找到出处；无跨时代混搭	小月	镜头-脚本对照表
音画同步	旁白与画面对应到 ±0.5 秒内；音乐在片尾自然收束，无突兀截断	小舟	时间线截图
术语统一	全片统一使用术语表用词（如"楼台"不写成"亭台"）	小江	术语表比对结果
史实与引文	诗句逐字与原文一致；涉及历史细节的表述标注核实方式	小江	核实记录
AI 参与标注	片尾与展示页注明哪些环节由 AI 生成、哪些由人完成；按标识要求添加显式标识	全组	标注稿
素材来源与授权	每个外部素材都有来源与授权方式记录；无未授权声音、肖像、字体	小舟	素材来源档案
技术说明与参数对照	八章流水线各环节的工具、关键参数、对照实验结论齐备	小舟	技术说明文档
文件与格式	命名符合规则；成片分辨率、帧率、响度符合展示要求；备份两份	小舟	交付包截图
无障碍与展示	字幕齐全、字号可读；音量在展示场地实测通过；时长控制在 2 分钟内	小月	现场测试记录

最后提醒一个心态问题。搭工作流的过程中，你一定会遇到"平台今天改了界面""昨天还能跑的流程今天报错"这类事——这不是你的错，而是这个领域的常态。本书的原则始终是：工具会变，方法与清单不会变。把注意力放在"这条流水线是否可复现、是否可验收、坏了怎么查"这三件事上，你换任何平台都能十分钟重新搭起来。

8.9 关键概念表

本章术语密度是全书最高的一章。把下面这张表当作"对账清单"：每一行都能用自己的话讲清楚，本章的原理目标就算达成；讲不清楚的，回到对应小节重读。

表 8-13　第 8 章关键概念（中英对照）

术语	英文	一句话解释
智能体	agent	以大语言模型为决策核心，能调用工具、保存记忆、规划步骤并在循环中推进任务的系统
四要素	LLM + tools + memory + planning	智能体的四个零件；缺任何一个都会退化成"只会说的助手"
ReAct 循环	ReAct	"想一步 → 做一步 → 看结果 → 再想"的循环，当前采用率最高的智能体推理范式
纯 ReAct / 包装式	plain / wrapped ReAct	裸循环会漂移；加规划器与监督器后长任务不偏，但成本 2–3 倍
工具调用 / 函数调用	tool use / function calling	模型输出结构化调用请求，由宿主程序执行，结果回填上下文
工具声明（模式）	tool schema	工具的"菜单条目"：名称、用途、参数与类型——写得含糊，模型就点错菜
模型上下文协议	Model Context Protocol, MCP	让任意模型客户端接上任意工具服务器的开放规范；2026-07-28 版改为无状态核心
无状态核心	stateless core	MCP 新规范的核心变化：握手与会话头被删除，协议版本随每个请求携带
多轮往返请求	multi round-trip requests, MRTR	服务器需要补充输入时返回"需要输入"的结果，客户端带上输入重试
检索增强生成	retrieval-augmented generation, RAG	先去资料里检索相关片段，再让模型"带着小抄"作答
切块	chunking	把长文档切成可检索的小块；约八成 RAG 失败发生在这里
向量库	vector store	存储文本块向量与元数据，供相似度检索使用
相似度检索	similarity search	用问题向量去找最相近的若干块（第 1 章余弦相似度的应用）
返回条数	top-k	一次取回多少条候选；调大不等于更好，还可能需要重排与去重
重排 / 融合排序	rerank / RRF	把候选按相关性重新排序、融合多路结果，再把上下文预算花在刀刃上
记忆	memory	短期（上下文窗口）、长期（跨会话偏好）与外部知识库三种形态
多智能体	multi-agent	多个角色协作完成一个任务；成本与延迟随之上升，未必比单智能体划算
层次式 / 点对点 / 竞争式	hierarchical / peer-to-peer / competitive	三种协作拓扑：主管分派、平级协商、互相辩论
角色路由	role-based model routing	规划用强模型、执行用便宜模型——最可靠的成本杠杆
编排	orchestration	谁在什么时机调用谁；编排做不好，能力再强也组不成流水线
停止条件	stopping condition	任务"算完成"的客观判据；"模型说自己完成"不构成判据
人在回路	human-in-the-loop	关键节点由人批准、监控或制定政策；本质是"有拦截权力的检查点"
验收清单	acceptance checklist	把"好不好"变成可判真假的条目，逐条留证据
审批门	approval gate	生成 → 暂存 → 审批 → 发布的关卡；一个审批人、一个动作
留痕	audit trail	保留 AI 原判与人类裁决两者，分歧可回溯
幂等	idempotent	同一写操作重复执行不产生额外后果——重试安全的前提
最小权限	least privilege	只给工具开必要的能力范围；删除、支付、发布必须人工确认
交互媒体	interactive media	观众的选择成为作品的输入，内容在观看过程中继续生成

8.10 交互媒体：当智能体走进作品内部

本章讲到这里，智能体一直是"制作工具"：它帮我们生成分镜、检索资料、串接流水线。但还差最后一步——智能体也可以走进作品内部，成为作品的一部分。这就是**交互媒体**：观众的选择成为输入，内容在观看过程中继续生长。这一节把它讲清楚，因为它是"智能体"与"媒体创作"真正的交汇点，也是你们期末作品可以选择的加分方向。

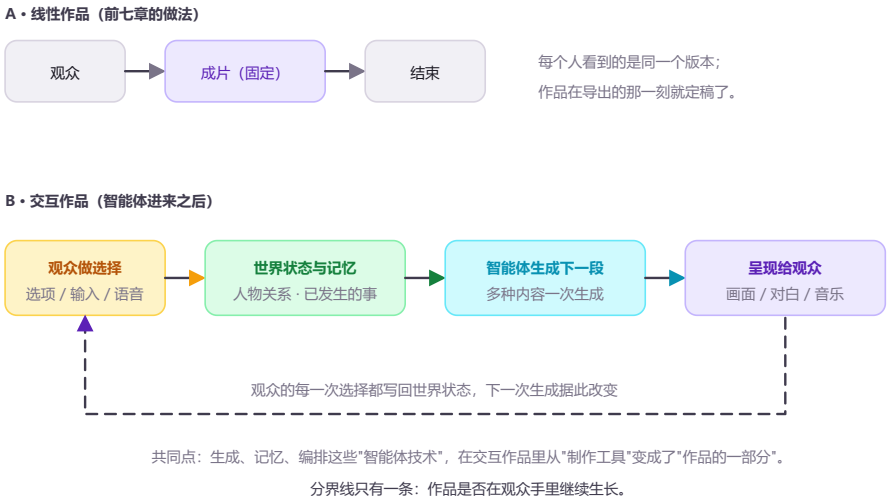


图 8-10　线性与交互：一条分界线。上半段是我们前七章做的线性作品；下半段把"世界状态 + 记忆 + 生成"放进作品内部。中间那条回环，就是"互动"这件事的技术本质。

◆ 技术前沿：小模型、实时与"共享记忆"

这个方向在 2025—2026 年从实验室走进了展会和商店，有几个进展对创作者特别有启发。

在学术与原型一侧，2026 年 7 月的 SIGGRAPH Appy Hour 上出现了一个叫 **Lost Spirits** 的 Godot 游戏原型：由一位艺术家单独完成，开发窗口两个月——他用旗舰模型做实时代码生成，用**本地运行的小模型**驱动 NPC，让非脚本对白由"情感状态机"调解出来，追求所谓"涌现叙事"。另一个例子 DefameLM 更进一步：只用**一个微调过的小模型**服务整个游戏循环，靠"重试直到成功"来应对实时延迟——这是"小模型 + 重试"路线的一个可教学范例。CVPR 2026 的一场工作坊则给出了视觉小说式冒险游戏的多智能体框架：实时协调语言模型、文生图与音乐生成，靠**共享记忆、动态故事图、角色知识库与情绪对齐音乐**维持一致性。

把这些原型的技术选择抽象出来，会发现三条共同的工程智慧：**① 用本地小模型换取低延迟（实时场景里，慢一步就出戏）；② 用"生成 + 重试"代替"一次生成到位"（把概率问题变成可接受的成功率问题）；③ 用共享记忆与故事图替多个模型"对齐世界观"**——第三条正是本章 8.5、8.6 两节知识在"作品内部"的重演。

◆ 行业图景：从"可以生成"到"可以玩"

商业与内容一侧的变化更快。2026 年上半年，国内主要平台几乎同时推出了 AI 互动内容产品，我们把公开报道里的事实整理成一张表——注意看每一行最后一列的"制作规模"，那是这门课最想让你看见的东西。

表 8-14　国内 AI 互动影游的一批公开实践（2025—2026 年报道）

机构	产品 / 平台	事实要点
字节	《诡舍》（2026 年 3 月）	AI 仿真人互动短剧：观众在关键剧情节点作出选择、影响后续发展，角色与场景全部由 AI 生成
	《不问凡尘》	已上线 Steam，首发折后 19.9 元，画面完全由 AI 视频模型生成，**团队不到 20 人**
	剪映 ICG Studio（2026 年 9 月发布/内测）	AI 互动影游创作平台：创作者定好主线与关键抉择点，**与智能体对话即可生成多条支线剧情与分镜**；工作流为"蓝本确认 → 剧本大纲 → 全局资产 → 分镜生成"
腾讯	造化工坊（2026 年 5 月）	AI 互动叙事创作与体验平台，报道称已积累**超过 570 款**互动影游
	探梦 DreamNow	AIGC 内容创作与展示平台，已测试互动影游《魏晋风骨》：观众选择剧情分支、作出人物抉择，解锁不同走向与结局
芒果 TV	《公主请自重》（2025 年 12 月）、《马上来财》（2026 年春节）	后者为 AIGC 互动影游，点击与滑动收集碎片解锁剧情分支，报道称总播放量破 1500 万、单日独立访客峰值超 33 万
拼多多	若干互动短剧（2026 年 3 月）	在电商场景内试水：用户选择分支决定走向，部分商品与道具可点击跳转购买
PixVerse	R2 世界模型与三款 AI 互动影游（2026 年 8 月报道）	采用"预置视频分支 + 实时 AI 生成"的混合模式，支持选择、文字、语音与即时反应输入；世界模型能记住玩家的修改，故事与角色随之调整
AltStory / 互影科技	AltFlow 创作平台 + 互动视频引擎	用智能体拆解小说与剧本、提取人物关系与叙事转折点并生成分支剧情；报道称可在 125 秒内完成 50 万字小说拆解，由 31 个专职 AI 智能体覆盖人设、场景到可视化互动剧本生成

行业规模也值得记一笔：**2026 年第一季度全行业上线微短剧约 12.8 万部，其中 AI 微短剧约 12.2 万部，占比超过 95%。**这个数字既说明产能被彻底改写，也说明导论里那个问题（当生成变得廉价，创作还剩下什么）在这一年有了新的答案：当"做出一版"的成本趋近于零，稀缺的就变成了"值得被做出来的那一版"。

> ⚠️ **必须写进去的另一面**
>
> 互动内容被视为微短剧的"新变量"：AI 降低了多线剧情的制作门槛，互动性也能提高完播率；商业化上，AI 互动影游偏向买断、会员或内置付费，AI 互动剧偏向广告变现。但报道与研究一致提醒：**AI 可以生成大量分支结局，却很难保证每个人物性格与因果链条都合理**——叙事深度与人物逻辑仍需人类把关。这句话值得抄在笔记本扉页上，它是本章所有技术乐观主义的刹车片。另有三项开放挑战被反复提及：**连贯性、延迟、长期 NPC 记忆。**你在期末作品里如果尝试互动，把这三点写进"技术说明"的"已知不足"一栏，会成为加分项。

> 📌 **五条趋势（写给你的未来三年）**
>
> 把这一节的观察收成五条趋势，它们会在你毕业前继续发酵：① 从**孤立生成内容**转向**集成可玩系统**；② **低延迟的本地/边缘小模型**让"实时生成"成为可能；③ **多智能体编排**成为保证连贯性的主要手段（编剧、角色设计师、场景设定师、画家分工 + 共享记忆）；④ AI 持续**降低单人与小团队的门槛**（单人两个月的原型、不到 20 人的成片团队、四人团队的商店页）；⑤ 开放挑战仍是**连贯性、延迟与长期记忆。**你的机会就藏在这五条的缝隙里。

8.11 案例教学：主线收官与三个可复用的样例

原理讲完了，这一节分两层看案例：先看你们身边的——春江小组怎么用一条三智能体工作流把前七章的全部资产串成一部成片，并完成课程史上第一次"公开验收"；再看三个外部样例，它们分别演示了**单人如何驾驭实时互动**、**大量素材如何被智能体拆解**、**创作平台如何把智能体藏进工作流**。

图 8-11　成片画面：《春江花月夜》收官镜头。八章的资产在智能体工作流里汇合，最终落在这样一帧上。它同时也是这门课的一个隐喻：画面可以极简，背后不能极简——每一处留白，都是流水线上一次明确的取舍。 AI 生成

◆ 案例 8-B　春江小组：三角色工作流串起前七章

案例 8-B　两分钟，一条流水线，三个人

背景。第 11 周，距离期末展示还有 9 天。春江小组手里的资产清单已经很长：小江定稿的脚本与旁白配音、小月画的分镜图与视觉设定、小舟从前面各章攒下的镜头片段、配乐与场景资产。三个人都会用各自的 AI 工具，但"谁会什么"和"东西在哪"只存在于群聊记录里——第一次联跑时，小舟光是确认"哪一版配乐是最终版"就花了四十分钟，还差点用错一版旧素材。

做法。小江提议：既然我们三个人分工这么清楚，那就不该只有一个"全能助手"，而应该把分工照抄成一条工作流。她画了一张三角色草图——**策划（小江出方案）→ 生成（小舟跑工具）→ 审核（小月按清单把关）**——然后把它搬上零代码平台。三个角色各自的职责、能碰的文件、必须遵守的规范，全部写进各自的提示词里；三个人类的角色则退后一步，变成"给策划角色定目标、给审核角色定清单、在争议时拍板"。

波折。第一次全片联跑，审核角色和生成角色"吵"了一下午：小月定的清单里有"全片时间线统一在同一时期"，但这条要求没有写进策划角色的意图描述，于是小月每指一次，小舟就要重跑一批镜头；到下午四点，同一个镜头已被重生成多次，额度报警。三个人停下来复盘，发现问题不在角色数量，而在**停止条件没定义**——工作流里没有任何地方写着"这一幕算完成"。他们补了三样东西：每幕的"完成的样子"（可判真假）、最多重试两次、超限升级给人；以及一个"写完即暂存、小江确认才进正片"的审批门。第二天重跑，一次通到底。

结果。第 11 周三的公开展示上，各小组的成片逐部放映，评审组对照验收清单打分。春江小组的短片完整放映、全程无事故；评审组给出的三条改进建议（"转场偏密""第二段的配乐盖住了旁白""结尾留白可以再长半拍"）被记进复盘笔记，成为下一届学生的"前人经验"。三个人在个人贡献说明里分别写清了各自负责的范围与 AI 介入程度——这份说明和作品一起归档，也是本章 8.14 节要展开的重要一课。

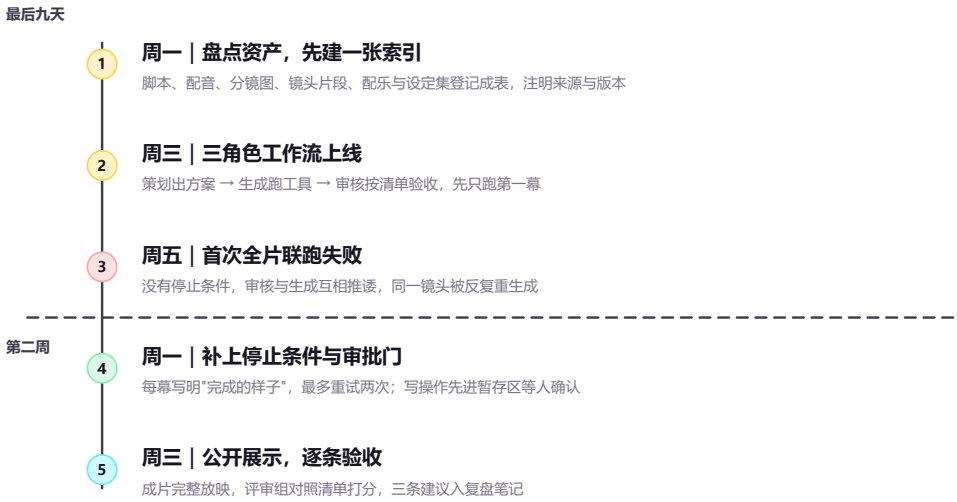


图 8-12　春江小组最后九天的推进时间线。注意第三与第四个节点之间的转折：真正让工作流"一次通到底"的，不是更强的模型，而是补上了停止条件与审批门。

> **💡 从这个案例里拿走三句话**
>
> ① **角色是照着人的分工抄的**——先有清楚的人类分工，才有清楚的多智能体分工；如果三个人自己都说不清谁管什么，交给智能体只会放大混乱。② **多智能体上线第一天的失败，几乎都不是能力问题，而是流程问题**——没有完成定义、没有重试上限、没有升级路径。③ **人退后，不等于人退出**——小江、小月、小舟从"执行者"变成"目标制定者、清单作者与拍板人"，这正是本章反复强调的人机分工。

◆ 外部样例：单人、大批量、平台化

案例 8-C　Lost Spirits：一个人，两个月，一间"会说话的屋子"

背景。交互作品常被默认为"团队项目"。但 2026 年 7 月的 SIGGRAPH Appy Hour 上，一个叫 Lost Spirits 的 Godot 原型给出了反例：它由一位艺术家单独完成，从开工到展示只有两个月。

做法。技术选择非常克制：旗舰模型负责"实时代码生成"（在开发过程中边写边让模型补代码），而真正让作品"活起来"的 NPC 对白，交给一个**本地运行的小模型**——这样延迟低、成本低，还能离线跑；非脚本对白由"情感状态机"调解，让角色的反应看起来从情绪里长出来，而不是从台词库里抽出来。作者把这种效果称为"涌现叙事"。

启示。三条。第一，**旗舰模型不一定全程在场**：开发期用强模型、运行期用本地小模型，是单人项目控制成本与延迟的常见组合，正对应 8.6 节讲的"角色路由"。第二，**"重试"是一种设计**：实时生成必然出错，把"生成 → 校验 → 重试"写进系统，比追求一次成功更现实。第三，**两个月做出原型靠的是范围控制**——一个人不可能同时做美术、程序、剧本，他做的是"把 AI 能力设计成作品机制"的那一部分，其余交给模型。这正是本章想教给你的取舍。

案例 8-D　AltFlow：31 个智能体，125 秒，50 万字

背景。互动叙事平台的瓶颈在"拆解"：一部 50 万字的小说要改成互动剧本，人工梳理人物关系与叙事转折点往往以周计。互影科技的 AltFlow 平台给出的答案是"把拆解本身做成流水线"。

做法。平台用智能体拆解小说与剧本、提取人物关系与叙事转折点并生成分支剧情；公开报道称，它由 **31 个专职 AI 智能体**覆盖从人设、场景到可视化互动剧本的生成，能在 **125 秒内完成 50 万字小说**的拆解。注意"专职"二字：这不是 31 个全能助手各干各的，而是 31 个各有窄职责、各有输入输出格式的岗位——和你把期末作品拆成子任务表（8.13 节实训一）是同一套方法，只是规模大成百上千倍。

启示。第一，**大规模拆解的前提是格式统一**：31 个岗位能互相接手工作，靠的一定是统一的中间产物（人设卡、关系图、转折点表），而不是自然语言的"聊天记录"。第二，**速度来自并行与专职，而不是来自更大的模型。**第三，请你对"125 秒"保持职业敏感：把 50 万字"拆开"是快的，把拆开的东西"判断对"是慢的——这一节所有案例的最后一句话都应该是"人把关"。

案例 8-E　剪映 ICG Studio：把智能体藏进创作工作流

背景。2026 年 9 月发布/内测的剪映 ICG Studio，是一个面向 AI 互动影游的创作平台。它对本章特别有价值，因为它演示了"平台如何替你把编排做好"。

做法。工作流是四步：**蓝本确认 → 剧本大纲 → 全局资产 → 分镜生成。**创作者先定好主线与关键抉择点，然后**与智能体对话，即可生成多条支线剧情和分镜**——你不需要自己搭循环、写工具声明、管上下文，平台把本章 8.1—8.5 的机制都封装在"蓝本"和"全局资产"这两个概念里：蓝本是给策划角色的约束，全局资产是把 8.5 节的知识库（人物设定、场景设定）接给所有生成角色。

启示。第一，**平台的抽象层次就是你的操作层次**：在 ICG 里你不操作"循环与工具"，你操作"蓝本、抉择点、资产"——但你得懂底下在发生什么，才知道出问题时该往哪查（这就是原理课的价值）。第二，**交互影游的门槛正在从"会用工具"变成"会设计抉择"**：分支写在哪里、每个抉择改变什么，是创作判断，不是技术问题。第三，能自动生成多条支线的工具，也最容易生成大量"用不上"的支线——验收清单（8.7 节表 8-10）在这里同样适用：每条支线都要回答"删掉它，作品会变差吗"。

8.12 失败案例诊所

本章原理部分已经埋了两个案例：F8-1"参数对了，单位错了"（工具调用层）和 F8-2"它引用了上一版设定"（知识库层）。这一节再集中问诊两个**智能体特有的病**——它们都不是"模型不够聪明"，而是"系统设计漏了一环"。

失败案例 F8-3　它说"已完成"，其实什么都没做

症状。某组把"批量整理素材并生成缩略图"交给助手后去上课。回来发现任务还在转，日志里排着二十多条记录，每一条的结尾都是"已修复！""问题已解决""继续优化中……"，看上去干劲十足；但打开输出文件夹——空的。助手循环了一下午，一步也没落地。

诊断。这是智能体最典型的两种病叠在一起：**自我肯定**（模型对自己的产出天然偏乐观，倾向于宣告成功）与**无事可做的空转**（反复"计划"和"总结"，却不调用任何工具，因为没有任何一条规则告诉它"必须动手才算数"）。追问之后发现三个设计缺口：① 完成判据来自模型自己的话——"你说完成就算完成"（正是 8.7 节讲的那个根问题）；② 工具返回的真实结果没有被检查——文件到底生成了没有，没人问；③ 循环没有上限、失败没有升级路径。

处方。四招，按性价比排序：**① 把完成条件外部化**——"输出文件夹里出现 12 张命名规范的缩略图"是可判真假的，写进任务描述；**② 每轮循环必须留下一个客观证据**（工具调用记录、文件清单），只有计划没有动作的轮次直接判为无效；**③ 设硬性上限**——最多循环 N 次、最多重试 2 次，超限立即停下来找人或换方案，别让它在沉默中烧额度；**④ 记住那个乘法**——单步成功率 0.95 时，20 步全对的概率只有约 36%（8.2 节的公式），所以长任务必须靠检查点续命，而不是靠模型"这次一定行"的信心。

失败案例 F8-4　两个智能体吵了三小时，可它们谁也没错

症状。某组搭了"生成 + 审核"双智能体工作流：生成角色交出一张分镜图，审核角色退回，理由"不符合竖屏规格"；生成角色重新生成，再次被退；如此往复十几轮，两边都在认真工作，进度为零。负责人把两边的对话打出来一看，哭笑不得——生成角色按 16:9 的规格表在做（那是它上一轮收到的说明），审核角色按 9:16 的验收清单在审（那是负责人新加的要求）。**两边都没错，错的是给它们的是两套事实。**

诊断。这是多智能体系统最隐蔽的失败模式：**接口漂移。**研究早就观察到这类现象的规律：给协作中的智能体持续输入"对抗性批评"时，批评的质量差异会带来约 10 倍的最终成本波动——审核角色的价值是真的，但前提是**它手里的尺子和生成角色的尺子是同一把。**本例中三个环节同时缺失：没有"单一事实来源"（规格表只存在于负责人的一次口头补充里）；审核标准没有写成可判定项（"不符合竖屏"没写清分辨率、比例、安全边距）；没有仲裁机制（两边僵住时无人拍板）。

处方。第一，**建一份全局规格表**（术语表 + 资产规格 + 命名规范），三个角色都要读它、改它要走审批——这是 8.5 节知识库思想在多智能体场景下的应用。第二，**把清单条款写成检查项**："是 9:16 吗？分辨率 ≥1080×1920 吗？字幕在安全区内吗？"——能打勾的条款才拦得住东西。第三，**设仲裁与升级路径**：同一分歧出现两次以上，自动停止重试并把双方证据摆给人看。第四，回头想想 8.6 节那笔账：如果连"生成 × 审核"两个角色都协调不好，就不要急着上五个角色——多加一个角色，就多一条要维护的接口。

8.13 动手实践：从拆任务到公开验收

本章的实践与你的期末作业是同一件事。四个实践按顺序做下来，你会得到四份可直接进档案的材料：一张**子任务分解表**、一条**三角色工作流**、一份**对照实验记录**、一张**互评评分表的结果。**它们加起来的体量不小，建议课堂做骨架、课后补细节。

🛠 实践一：把"完成一部短片"拆成可验收的子任务（50 分钟 • 课堂）

基础

拿一张纸（或表格软件），以你们组的期末作品为目标，做两级分解。注意两条铁律：**每个最末级的子任务都必须能用"是 / 否"判断完成**；**每个子任务都必须写出负责人**（人或某个智能体角色，但最终负责到底的一定是人）。

1. **先写"目标"。**在纸的最上方写下不可直接交付的目标（"完成一部两分钟的短片"）——它是树根，不是任务。
2. **分三条线。**按内容（脚本、分镜）、声音（配音、配乐）、影像（镜头、剪辑）先把大块分开——这就是图 8-13 的第一层。
3. **拆到叶子。**继续往下拆，直到每个叶子都是"某人做某事、产出某物、可被判真假"。拆不到可判真假，说明还没拆完。
4. **给每个叶子配验收标准。**参考本章表 8-10 的十条清单；至少写清三件事——产出物长什么样、合格线是什么、谁验收。
5. **标出人的检查点。**在分解表上把"必须由人确认才继续"的节点圈出来（通常是：定稿类、发布类、涉及素材授权的），数一数有几个——少于两个，说明你的检查点太少了。

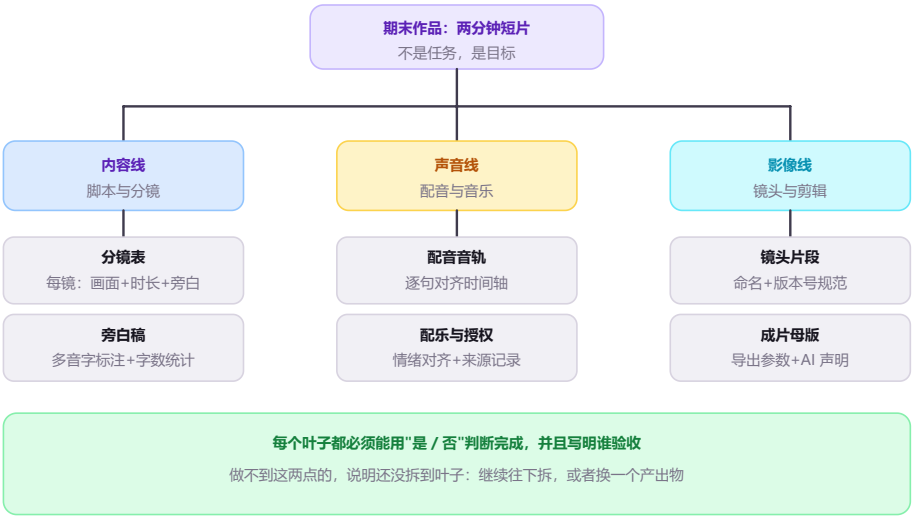


图 8-13　目标 → 三条线 → 可验收的叶子。树根是不可交付的目标；叶子是可判真假的子任务。中间层（三条线）可以按你们组的分工调整，但"叶子必须可验收"这一条不能改。

🛠 实践二：搭一条三角色工作流（60 分钟 • 课堂 + 课后）

进阶

在零代码平台上（扣子 / Dify 或你熟悉的等效工具，参见 8.8 节）搭一条最小可用的三角色工作流：**策划 → 生成 → 审核**。先用实践一的子任务表里最小的一个叶子做输入——例如"为第一幕生成三张候选分镜图"。

1. **建全局规格表。**把术语表与资产规格（画幅、分辨率、命名规范）写成一份文档，作为三个角色共同的参考（可直接用知识库功能挂上去）。这一步不做，你大概率会复现 F8-4。
2. **配策划角色。**职责：把目标翻译成方案与提示词，不接触工具、不碰文件。按下面的提示词卡配置。
3. **配生成角色。**职责：按方案调用工具（图片生成、视频生成、配音等），只负责执行与回报结果，不做审美判断。
4. **配审核角色。**职责：逐条对照验收清单检查产出，输出"通过 / 退回 + 具体理由"，退回理由必须指向清单条款。
5. **加护栏。**同一任务最多重试 2 次；超限升级给你；写操作（覆盖、删除、发布）一律先进暂存区。
6. **跑通一幕并计时。**记录：三轮角色各自耗时、总耗时、重试次数、你的介入次数。这份原始记录就是实践三的对照数据之一。

提交物：工作流截图 + 三个角色的配置说明 + 一次完整运行的记录。若平台支持导出，把工作流定义一并归档。

角色提示词 • 策划

✎ 策略层 · 不碰工具

你是短片《**作品名**》的策划角色。你的唯一职责是把导演的目标翻译成**可执行的方案**，交给生成角色。

- 每次接到目标，先输出：这一步要什么、判断标准是什么、失败的常见原因是什么。
- 方案必须包含可直接使用的提示词或参数，不写空话。
- 你不调用任何工具、不修改任何文件；需要素材时，明确写清向谁要什么。
- 你不判断作品"好不好看"，那是审核角色与导演的事。

全局规格 见《术语表与资产规格》（必读）　输出格式 目标 / 方案 / 提示词 / 交办对象　禁忌 不调用工具、不宣布完成

角色提示词 · 生成

✎ 执行层 · 只动工具

你是《**作品名**》的生成角色。你只做一件事：按策划角色给的方案调用工具，如实回报结果。

- 每次执行前复述收到的参数；参数缺失或矛盾时**停下来问**，不要自己猜。
- 执行后如实上报：调用了什么工具、用了什么参数、产出文件在哪、成功还是失败。
- 失败时报告原始错误信息，不要用"已修复""已优化"之类的说法掩盖失败。
- 你不做审美判断，不自行修改方案；重复失败两次即停止并升级。

可用工具 图片生成 / 视频生成 / 配音 / 素材检索（按平台实际情况）　命名规范 见《术语表与资产规格》　重试上限 2 次，超限升级

角色提示词 · 审核

✎ 质量层 · 只对照清单

你是《**作品名**》的审核角色。你手里只有一把尺子：**验收清单**。逐条检查，逐条给出结论。

- 对每条清单输出：通过 / 不通过 + 证据（看哪里得出的结论）。
- "不通过"必须指向具体条款与具体位置，禁止只说"感觉不对"。
- 你不修改作品，也不重新生成；你只裁决，并把退回理由写清楚交给生成角色。
- 清单没覆盖到、但你认为有问题的地方，单独列"清单外意见"，交由导演决定是否补进清单。

验收清单 见本章表 8-10 与本组补充条款　输出格式 条款号 / 结论 / 证据 / 退回理由　仲裁 同一分歧两次以上 → 停止重试，升级给人

🛠 实践三：单智能体 vs 多智能体——一次受控对照（50 分钟 • 课堂 + 课后）

进阶

把实践一中**同一个子任务**（如"为第一幕生成三张候选分镜图并选定一张"）用两种方式各跑一遍，其余条件全部固定——同一目标、同一组工具、同一张验收清单、同一批参考素材。这就是"多一个变量"的受控实验：唯一的变量是**智能体数量**。

- **A 版（单智能体）**：一个助手从头做到尾，一句话交付全部要求。
- **B 版（多智能体）**：用实践二的三角色工作流做同一件事。
- **两边都用同一张验收清单打分**（逐条通过率 + 你们组给出的 1–5 分主观分），并记录成本（额度消耗或费用估算）、总耗时、你的介入次数与重试次数。

提交物：填写完整的表 8-15 + 一段 150 字以内的结论（不超过 150 字，写清"这次实验后，我们组决定在期末作品里用几个角色、为什么"）。

挑战

把你的实测数据与公开研究对读（本章表 8-9）：经济学元分析给出多智能体约 **30 倍 token 消耗、28–35 倍成本**；Chalmers 的实验里单片式 81.1% / 0.083 美元，层次式 54.4% / 0.308 美元；AgentRadio 用 6.6 倍成本换来 62.1% 对 57.2% 的得分。回答一个问题："我们组的实测与这些数字差多少？差距更可能来自哪里——任务类型、角色设计，还是验收标准？"把答案写进对照记录，作为你们"成本-质量"决策的依据。

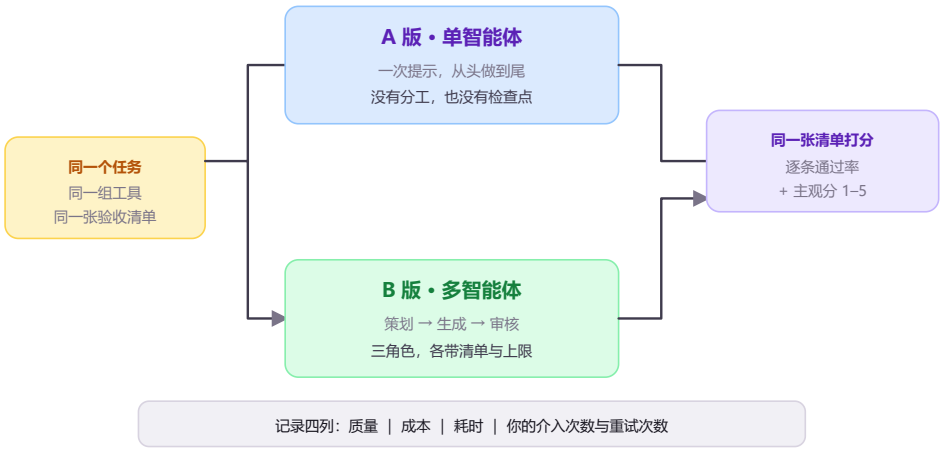


图 8-14　受控对照：只让"角色数量"这一个变量动。左边和右边跑的是同一个任务、同一组工具、同一张清单；唯一不同是 A 版一个智能体、B 版三个。学不会"只动一个变量"，你拿到的数据就没法解释。

表 8-15　单 / 多智能体对照实验记录表（模板，复制到你的创作档案）

记录项	A 版（单智能体）	B 版（多智能体）	差值 / 备注
验收清单通过率	– / 10	– / 10	逐条记录，不通过项写理由
主观质量分（1–5）	–	–	同一位组员评两次，避免两套标准
成本（额度 / 费用估算）	–	–	与表 8-9 的 28–35 倍量级对照
总耗时	–	–	含等待与重跑时间
人类介入次数	–	–	你出手干预、修改、拍板的次数
重试 / 循环次数	–	–	超限是否触发了升级
结论（≤150 字）	这次实验后，我们组决定在期末作品里用 ___ 个角色，理由是……		

🛠 实践四：公开展示与互评（一次课 · 课堂）

基础

期末展示不是"把片子放一遍"，而是一次**验收行为**——本章 8.7 节的整套流程，这一次由全班一起执行。

1. **提交材料包（放映前 24 小时）。**每组提交：成片文件、AI 参与程度声明、个人贡献说明、工作流截图、对照实验记录（表 8-15 与实践一的分解表）。缺一项，展示顺延——流程严肃，作品才被认真对待。
2. **放映。**每组 3 分钟成片 + 2 分钟讲解（讲什么：你们的工作流长什么样、哪一步最难、你们做的最关键的一个人工判断是什么）+ 2 分钟问答。
3. **互评。**其余各组用表 8-16 打分并写一条"值得学习之处"与一条"如果是我会怎么改"。评分表写明打分依据，不允许只给分数。
4. **教师与评审组合议。**对照表 8-10 的十条清单逐组核验文档是否齐备，结合互评平均分给出最终评价与三条改进建议。
5. **写复盘笔记。**每组把收到的建议整理成"复盘笔记"归档，留给下一届——这是本课程的传统，也是"公开验收"能持续变好的秘诀。

表 8-16　期末作品公开展示互评表（模板，每个评分项写一句依据）

评分维度	观察什么	分值	得分与依据
叙事与完成度	两分钟里讲清了一件事吗？起承转合完整吗？结尾站得住吗？	25	–
视觉与声音质量	画面风格是否统一？配音、音乐与画面的对位是否成立？有无明显瑕疵（闪烁、断音、时长错位）？	25	–
智能体使用与可控性	工作流设计合理吗？角色分工是否清楚？有没有可见的停止条件、审批门与留痕？人对产出负责的痕迹清楚吗？	25	–
文档与诚信	AI 参与程度声明、个人贡献说明、来源与授权档案是否齐备、真实、可核查？	15	–
展示与问答	讲清楚了最关键的一个人工判断吗？对提问的回答是否诚实（包括承认不足）？	10	–
值得学习之处	一句话，写给下一个小组	–	–
如果是我会怎么改	一句话，必须具体（指向某个镜头、某段音乐或某个流程）	–	–

8.14 课程思政：判断在人，责任在人

本章技术内容的每一次"升级"，其实都在把同一个问题推到你面前：当系统越来越会"自己办事"，**谁在负责**？这一节把它分成两层来谈——一层关于责任，一层关于创作身份。它们是全书唯一没有"技术替代方案"的两件事。

🇨🇳 思政一　人在回路：把"检查点"变成一种品格

为什么必须讲这个。本章给了你很多"放手"的理由：让智能体自己循环、自己检索、自己审核。但所有公开的工程经验都在提醒同一个事实——系统越自主，"人对哪里负责"就越要写清楚。2026 年 2 月的一篇论文警告：超过 40% 的智能体项目预计到 2027 年会因"编排鸿沟"而失败——不是模型不行，而是人把流程交给系统之后，忘了问"出错时谁知道、谁能停、谁担责"。监管的时间表也在收紧：欧盟的 AI 透明度义务自 2026 年 8 月 2 日起适用，高风险系统的人类监督义务到 2027 年 12 月生效。规则的语言很枯燥，它说的其实就是三个字：**要有人。**

在本课程里，这意味着什么。你的期末作品里，必须能找到"人的手"。它不是一句口号，而是三样看得见的东西：① **检查点**——至少两个"不通过就不往下走"的关卡，通常设在定稿与发布之前；② **留痕**——保留 AI 的原判与你的裁决两者，尤其保留你自己否掉 AI 建议的那些记录（那些记录最能证明你在创作，而不是在收件）；③ **署名与担责**——作品出问题时，不能说"这是 AI 生成的"。专业领域早就想清楚了这件事：医生可以用 AI 辅助读片，但诊断书上是医生签名；工程师可以用软件计算，但图纸图签上是他自己的名字。工具承担计算，人承担判断——**责任不能外包给工具，就像它不能被转交给一个不在场的合作者。**

写给你的长期提醒。你未来使用的工具会比今天更强，"放手"的诱惑会更大。请记住本章那个乘法：单步 95% 可靠、走 20 步就只剩三分之一多一点的胜算。**回路里那只手不是对 AI 的不信任，而是对整个系统的常识**——就像开车系安全带，不是因为不信任自己，而是因为知道意外是概率的一部分。

🇨🇳 思政二　创作主体性与透明协作：把"我做了什么"如实写下来

为什么必须讲这个。第 8 章是全书 AI 成分最高的一章：脚本可以由模型起草、画面可以由模型生成、工作流可以由平台编排。工具越强，"哪些是我的创作"这个问题的答案就越模糊——而模糊之处，正是学术诚信最容易失守的地方。规则层面，国家已给出清晰的方向：《互联网信息服务深度合成管理规定》要求对合成人声、人脸等可能导致公众混淆误认的内容在合理位置**显著标识**；《人工智能生成合成内容标识办法》进一步区分**显式标识**（显著位置提示 AI 生成属性）与**隐式标识**（数字水印等技术，用于溯源与监管核查）。也就是说，"标注"在这门课上不只是道德要求，它已经是有法可依的操作规范。

在本课程里，这意味着什么。三条硬性要求，请当成作品的"格式条款"：

- **① AI 参与程度声明。**每个作品附一页说明：哪些环节由 AI 生成（并注明工具与日期）、哪些环节由对话助手起草后经人重写、哪些完全由人完成。写清楚"生成"与"选定/改写"的区别——后两者往往才是创作发生的地方。
- **② 个人贡献说明。**每位组员写清自己负责的范围、完成的工作量、参与的关键判断（例如"否掉了哪一版方案、为什么"）。这门课的期末作品必须注明每个人的贡献——这不是走流程，而是让"署名"重新变得有意义：你签的是你真的做过的那部分。
- **③ 学术诚信红线。**杜绝代写（让他人或其他 AI 冒名完成应属本人的工作）、挂名（参与说明与实际贡献不符）、抄袭与挪用（未标注来源的素材、未授权的音乐与形象）。本章 8.8 节的素材来源档案正是为此准备：**可核查，是诚信最好的姿态。**

一个更高一层的判断。有人担心"用了 AI，创作者的身份就贬值了"。我们的判断是相反的：当生成变得廉价，**可识别的人类判断反而升值。**观众与评审真正能记住的，不是你用了什么工具，而是你在这里做了什么选择——为什么是这个结尾而不是那个，为什么删掉那段漂亮的镜头，为什么守住了一个没人会注意到的细节。技术会换代，这些选择不会。透明地标注 AI 的参与，不会削弱你的署名；恰恰相反，它让你的署名更可被信任。

8.15 回望全书：一条流水线的八个站点

这是本书的最后一节。此刻你手里应该有一部能放给全班看的短片，一部自己拆解过的子任务表，一条跑通过的智能体工作流，以及一份诚实的对照实验记录。把镜头拉远——从第 1 章到现在，你其实一直在修建同一条流水线，只是每一个月你都在往上面接一个新站点。

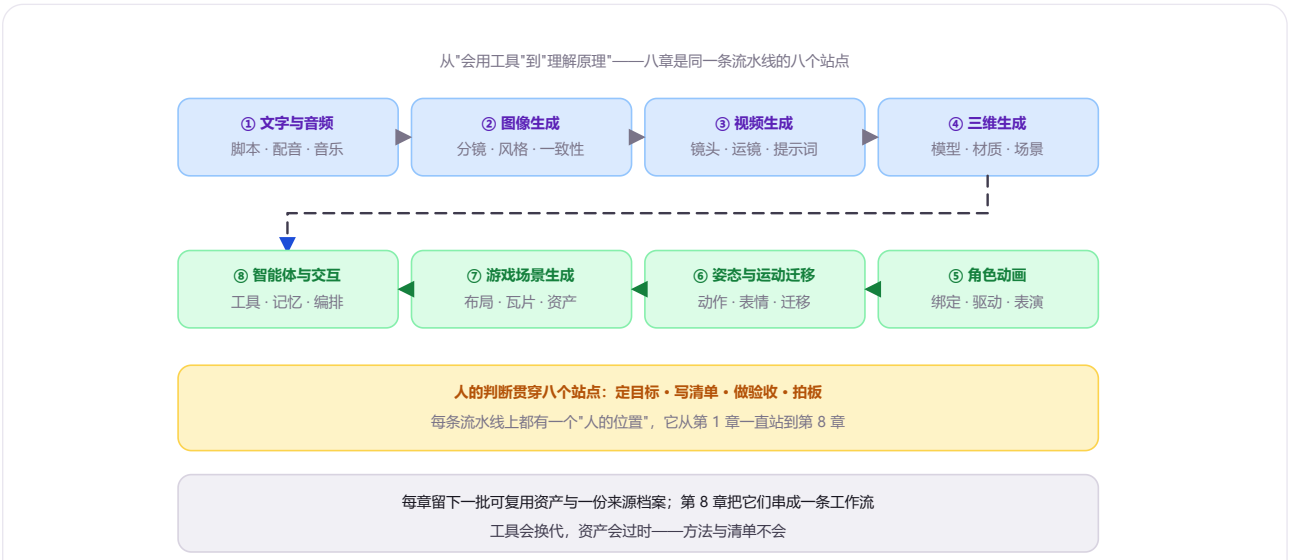


图 8-15　八章流水线：从素材到成片。读法有两种：顺着①到⑧看，是"一个作品是怎么被做出来的"；沿最下面两条带子看，是"什么一直没有变"——人的判断，与方法。

◆ 四种"长大"

从"会用工具"到"理解原理"。第 1 章你学的是"怎么让工具说对话"；到第 8 章，你要能回答"它为什么会那样做、在哪里会失效、我该在什么地方接住它"。前者让你能用今天的工具，后者让你在工具换代后不必重新开始——因为你要的是原理，而原理这一层，变化最慢。

从"生成一件东西"到"经营一条流水线"。初期你的成就感来自"这一张图真好"；现在你知道，一张好图只是流水线上的一个中间产物。作品的质量取决于最弱的那一环，而不是最强的那一环——这是把作业做成作品、把创作变成可重复实践的关键一步。

从"我让 AI 做"到"我知道哪里必须是我"。这一路上你做过很多次这样的判断：哪句话必须是自己的，哪个镜头必须自己盯，哪个"通过"必须自己说。判断力不是从知识里长出来的，而是从"按这个键之前先想一秒"的习惯里长出来的。

从"做完就行"到"可以拿给别人看"。验收、留痕、标注、复盘——这些在初期看起来像麻烦事的动作，最后构成了一个创作者的专业性。你能坦然展示的不只是作品，还有作品背后的过程；这大概就是这门课能给你的、最不靠工具的一项能力。

◆ 交还给你的一句话

八章讲完，其实只讲了一件事：AI 是一种力量极强但**需要被安放**的能力——安放在正确的分工里（智能体与多智能体），安放在可靠的知识里（检索与记忆），安放在明确的验收里（清单与停止条件），最后，安放在一个愿意为结果负责的人手里（也就是你）。人在回路里的"回路"两个字，说的不只是流程，也是责任绕不开的那个圈。

现在，把片子放给全班看吧。灯光暗下来的时候，有一件事已经确定：这条流水线上每一个"通过"，都是你自己按下的。祝你的下一部作品，也这样开始。

◆ 把这一章压缩成七句话

1. 智能体 = 大语言模型 + 工具 + 记忆 + 规划；缺一个零件，它就会退化成"只会说的助手"——四个零件都可以用"缺什么补什么"的方式检查。
2. ReAct 的直觉是"想一步、做一步、看结果、再想"：它让模型不再一次性猜答案，而是每一步都拿到环境的反馈；当前绝大多数智能体都建在这个循环上。
3. 工具调用的本质是"模型输出结构化指令 → 程序执行 → 结果回填"：模型从不真正"动手"，它只会写调用请求；参数写得含糊，它就点错菜。
4. MCP 是给工具装的"统一插座"：一次声明，处处可用；2026-07-28 版改为无状态核心并加入能力扩展与授权机制，但它仍在快速演化，锁定与返工风险要算进项目。
5. RAG 的两条流水线里，约八成失败来自入库侧（收集、清洗、切块）；记忆的关键不是"存得多"，而是"每次把对的那几块塞进上下文"。
6. 多智能体是一次要用数字换来的交易：分工、评审、辩论确实能提升质量与鲁棒性，但成本常是单智能体的几倍到几十倍——先问"这笔账划不划算"，再问"怎么分工"。
7. 停止条件必须外部化：验收清单可判真假、检查点有人有权力、重试有上限、超限能升级；"让 AI 自己说完成"从来不是完成标准，责任始终在人。

◆ 自测

❓ 自测 1：一位同学说"我用了一个特别强的模型，所以它就是智能体了"。请用四要素模型指出这句话的问题，并各举一个"缺零件"的症状。

答案：模型强只解决了四要素里的第一个（决策核心），不等于智能体。缺**工具**的症状：它能给你详细的修图方案，却做不出一个像素的修改（"只会说，不会做"）。缺**记忆**的症状：你对它说的"我们组的女主角叫阿澈"，换一段对话它就不认识了。缺**规划**的症状：多步任务做到一半跑偏、重复劳动、忘记最终目标。检查方法：拿任务清单逐条问"这个环节靠哪个零件"，缺哪个补哪个——补的顺序通常是先工具、再记忆、最后规划。

❓ 自测 2：用"想一步、做一步、看结果"解释：为什么 ReAct 式的智能体比"一次性让模型输出最终答案"更可靠？它又带来了什么新代价？

答案：一次性输出时，模型必须在没有反馈的情况下连续猜对每一步，错误会一路累积；ReAct 把任务拆成循环，每一轮"观察"都是一次纠错机会——上一步错了，下一步能看见并修正，因此长任务的成功率显著提高。新代价有三个：① **步数放大失败概率**（单步 0.95、20 步全对只剩约 36%，所以必须设检查点与上限）；② **成本和延迟随步数增长**（研究观察到包装式 ReAct 约为单次调用成本的 2–3 倍，步骤可达 20 步以上）；③ **新的失败模式**：死循环、自我肯定、工具误用——这正是本章失败案例诊所 F8-3 的主题。

❓ 自测 3：把"API 像点菜窗口"这个类比补完整：菜单、点菜单、传菜口分别对应工具调用里的什么？为什么说"菜单写得好不好"决定了智能体可靠不可靠？

答案：菜单 = 工具声明（名称、用途、参数与类型，即 tool schema）；点菜单 = 模型输出的结构化调用请求（含参数值）；传菜口 = 宿主程序执行后把结果回填给模型的环节。菜单写得含糊，模型就会点错菜：参数名语义不清（"time"到底是秒还是毫秒）、缺少取值范围、描述里没写"什么时候该用这个工具"，都会导致调用错误或该调用时不调用。这就是本章"工具设计六原则"（表 8-4）的由来——**工具说明书写给模型看，它和提示词一样重要**。F8-1 的"参数对了、单位错了"正是菜单没写清单位造成的。

❓ 自测 4：有同学说"我们接了 MCP，以后所有工具都能即插即用了"。这句话哪些部分对、哪些部分过于乐观？

答案：对的部分：MCP 确实把"每个工具各写一遍对接代码"变成"一次声明、多处复用"，生态增长很快（2026 年报道的 SDK 月下载量达数亿次、服务器数量上千），工具接入的边际成本显著下降。过于乐观的部分：① 协议本身仍在快速演化——2026-07-28 版做了无状态核心等大幅重构，并且有明确的弃用窗口与路线图，今天的接法明年可能需要返工；② 授权与安全仍是工程问题（生产级 OAuth 部署、客户端预注册等），规范并未替你解决；③ 存在"锁定"风险，业界对"某个厂商的原生做法"与"通用规范"之间的分歧有公开讨论；④ MCP 之于智能体是"接入标准"，不是"能力本身"——它不保证工具可靠、也不保证模型会用。合理态度：积极采用，但把返工成本写进计划。

❓ 自测 5：期末联跑时，助手为了回答"楼台材质怎么描述"检索到了三条互相矛盾的设定。请按"查病历"的顺序说明你会先查哪里、为什么。

答案：先查**入库侧**，再查检索侧，最后才是模型。理由：公开数据显示约八成 RAG 失败发生在收集、清洗与切块环节——优先排查性价比最高。具体顺序：① 源头有没有多个版本同时入库（F8-2 的故事：v1–v5 全在里面，谁都被召回得到）；② 切块是否把"这一版作废"之类的限定语切掉了，导致块自相矛盾；③ 元数据是否带版本/日期，检索时能否只取最新版；④ 相似度检索是否取回太多旧块（调小返回条数或加重排）；⑤ 最后才考虑拼装与提示——上下文里塞进了两条矛盾信息，模型只能二选一猜。处方：入库前做版本标记与去重，检索时按版本过滤，并把"以哪一版为准"写进系统提示。

❓ 自测 6：某组想给期末作品"上强度"：把工作流从三个角色扩到七个，每人一个角色。请你用本章的成本账给出劝告，并提出一个替代方案。

答案：劝告要点：多智能体的收益不是免费的——经济学元分析给出约 30 倍 token 消耗与 28–35 倍成本；Chalmers 实验里层次式多智能体成本 0.308 美元、结果 54.4%，而单片式 0.083 美元、结果 81.1%；即便成功的例子（如 AgentRadio）也是用 6.6 倍成本换 4.9 个百分点；而且每多一个角色，就多一组需要维护的接口与一处协调失败的可能（F8-4）。替代方案（任选其一并说明理由）：① 维持三角色，把省下的预算投到"角色路由"——策划用强模型、执行用便宜模型，实测最可靠的成本杠杆（公开数据显示角色路由可降低约 57% 成本）；② 先在单智能体上加一张验收清单与两个检查点，用对照实验证明"确实不够用"再扩编；③ 把"七个角色"降级为"三角色 + 两个按需调用的专职工具"。

◆ 拓展阅读

- 《Model Context Protocol 规范》（2026-07-28 版）与官方服务器目录——不必通读，重点看"无状态核心""能力扩展"与"授权"三节；配合本章表 8-5 阅读，理解"协议进化"这件事的节奏感。
- 一份公开的多智能体成本研究（本章表 8-9 任一来源，如斯坦福的等预算对照实验或经济学元分析）——读它的实验设计部分：变量怎么控制、结论的边界在哪里。这比记住倍数更有用。
- Lost Spirits（SIGGRAPH Appy Hour 2026）与 CVPR 2026 工作坊的互动叙事论文——看一位创作者如何把"小模型 + 重试 + 情感状态机"组合成作品机制，是本章 8.10 节最好的延伸。
- 《互联网信息服务深度合成管理规定》与《人工智能生成合成内容标识办法》——通读"标识要求"部分，并在发布期末作品前逐条对照（附录 E 有操作清单）；这是本章思政二的法规依据。
- 张若虚《春江花月夜》全诗与历代评注——贯穿全书的主线原典；最后一章回望时，建议把全诗再读一遍，看看"江畔何人初见月"这一问，与你这一学期对"人在回路"的理解是否有了新的呼应。

附录 A–F

随查手册：术语 • 工具 • 提示词 • 评分 • 合规 • 选学

六份手册，覆盖"写得对、查得到、做得出、交得掉"的全部装备。

本附录怎么用

1. **平时常翻：**附录 A（术语）随章节查、附录 C（提示词库）随时复制。
2. **动手之前：**附录 B（工具速查）用来选型，附录 E（合规清单）用来"发布前三查"。
3. **期末冲刺：**附录 D（作品指南与评分表）逐条对照自查（学习目标：把评分标准当作验收清单）。
4. **学有余力：**附录 F（选学）通向"最小 API 调用"与本地工作流。

全书配套　工具与条款以官方页面为准

> 本附录是"工具书"性质的材料：不需要通读，需要在正确的时刻翻到正确的一页。所有工具名录与额度信息为编写本书时（截至 2026 年 10 月）检索所得，仅供参考；**使用前请以各平台官方页面的当前说明为准**——工具的世界里，唯一不变的是变化本身。

附录 A 术语表：中英对照速查

按主题分组，共约 120 条。建议的用法：每学完一章，回到对应分组"过一遍"——能用自己的话解释每一行，该章的原理目标就达成了。

◆ A-1　通用与人工智能基础（第 1 章）

表 A-1　通用与 AI 基础术语

中文	英文	一句话解释
模态 / 多模态	modality / multimodal	信息形态（文本、图像、音频…）/ 把多种形态统一处理
词元	token	模型处理文本的最小切块单位
分词器	tokenizer	决定文本如何切成词元的规则或程序
嵌入 / 向量表示	embedding	把内容映射成向量，语义近则距离近
潜空间	latent space	压缩后的语义空间；语义变换≈方向移动
余弦相似度	cosine similarity	用向量夹角衡量相似程度
大语言模型	LLM	以预测下一个词元为核心训练的大模型
Transformer / 注意力	Transformer / attention	主流架构 / 给上下文分配权重的机制
预训练 / 微调 / 对齐	pretraining / fine-tuning / alignment	学知识 / 学任务 / 学行为偏好
上下文窗口	context window	一次能读入的词元上限；宣称窗口常大于有效窗口
推理模型	reasoning model	先写长草稿（思维链）再作答的模型
思维链	chain-of-thought, CoT	让模型显式展开推理步骤的方法
提示词 / 负向提示词	prompt / negative prompt	输入指令 / 明确"不要什么"
零样本 / 少样本	zero-shot / few-shot	不给示例 / 给 1–3 个示例的提示模式
温度 / 核取样	temperature / top-p	控制取样随机性的两个参数
种子	seed	取样的随机起点，用于复现结果
幻觉	hallucination	流畅自信但不符合事实或凭空编造的输出
过度顺从	sycophancy	模型倾向附和用户观点、牺牲真实性的行为
检索增强生成	RAG	先检索资料再生成，降低幻觉

◆ A-2　声音（第 1 章）

表 A-2　声音术语

中文	英文	一句话解释
采样 / 波形 / 频谱	sampling / waveform / spectrum	逐点记录 / 时域曲线 / 频率构成
语音合成	TTS	文本→语音；分文本前端、声学模型、声码器三段
声码器	vocoder	把声学特征还原成波形
语音克隆	voice cloning	用少量参考音频复刻音色；零样本已是主流
音色 / 韵律	timbre / prosody	"是谁在说" / "怎么说"（语调、停顿、轻重）
拟音	foley	用真实道具为画面"演"出音效的工种
响度	loudness (LUFS)	整体音量感受，旁白与配乐需"配平"

◆ A-3　图像（第 2 章）

表 A-3　图像生成术语

中文	英文	一句话解释
扩散模型	diffusion model	学会"去噪"，从而能从噪声生成图像
加噪 / 去噪	noise / denoising	训练加噪学去噪；生成从纯噪声逐步去噪
采样步数	sampling steps	去噪迭代次数；越多越细，但收益递减
无分类器引导	CFG	控制"多听提示词"的强度参数
图生图 / 重绘强度	img2img / denoise strength	以图为起点生成 / 与输入图的差异程度
局部重绘 / 扩图	inpainting / outpainting	只改指定区域 / 向画面外扩展
条件控制	ControlNet 族	用线稿/深度/姿态等约束构图
流匹配	flow matching	2024 年后的主流训练范式，路径更"直"

◆ A-4　视频（第 3 章）

表 A-4　视频生成术语

中文	英文	一句话解释
文生视频 / 图生视频	text-to-video / image-to-video	从文字/图片出发生成视频
首尾帧	first / last frame	指定起止画面，中间由模型补全
时间一致性	temporal consistency	帧与帧之间不闪烁、不形变
运镜	camera movement	推、拉、摇、移、跟、升降等镜头运动
景别	shot size	远、全、中、近、特的取景范围
对口型	lip sync	让人物口型与配音对齐

◆ A-5　三维与动画（第 4、5 章）

表 A-5　三维与动画术语

中文	英文	一句话解释
网格 / 点云 / 体素	mesh / point cloud / voxel	三种显式三维表示：面、点、格
隐式场 / SDF	implicit field / signed distance field	用"每个点的距离值"隐式描述形状
神经辐射场	NeRF	用神经网络记住"从各方向看各点的颜色"
三维高斯泼溅	3D Gaussian Splatting	用数百万"发光椭球"拼出可实时渲染的场景
SDS 蒸馏	score distillation sampling	让 2D 生成模型当"老师"，指导 3D 形状优化
拓扑 / 面数	topology / poly count	面的连接结构 / 面片总数（性能关键）
UV 展开 / 法线贴图	UV unwrapping / normal map	把三维表面"摊平"贴图 / 用贴图伪造凹凸细节
PBR 材质	physically based rendering	按物理规律描述表面（颜色、粗糙度、金属度…）
骨骼 / 蒙皮 / 权重	armature / skinning / weights	骨架 / 把模型绑上骨架 / 每点受哪些骨骼影响
关键帧 / 插值	keyframe / interpolation	手工定"关键时刻" / 机器填补中间过程
IK / FK	inverse / forward kinematics	从"手放哪"反推关节角 / 从关节角正推位置
动作重定向	motion retargeting	把动作从一个骨架"翻译"到另一个骨架
自动绑定	auto-rigging	算法自动为模型生成骨架与蒙皮

◆ A-6　姿态与场景（第 6、7 章）

表 A-6　姿态与场景术语

中文	英文	一句话解释
姿态估计	pose estimation	从图像/视频中找出人体关键点的位置
关键点	keypoint	肩、肘、腕等关节位置坐标
置信度	confidence	模型对该关键点位置的把握程度
遮挡	occlusion	身体部位被挡住，导致估计困难
时序平滑	temporal smoothing	抑制逐帧抖动，让动作更稳定
身体模型	SMPL	参数化的人体三维模型（形状+姿态参数）
程序化生成	procedural generation	用算法/规则批量生成内容
天空盒	skybox	包裹场景的"天空贴图"球/盒
面数预算 / LOD	poly budget / level of detail	性能约束 / 远近距离用不同精度模型

◆ A-7　智能体与合规（第 8 章与附录 E）

表 A-7　智能体与合规术语

中文	英文	一句话解释
智能体	agent	LLM + 工具 + 记忆 + 规划的"办事系统"
工具调用 / 函数调用	tool use / function calling	模型输出结构化指令，程序代为执行
模型上下文协议	MCP	统一"模型接工具"的接口标准
多智能体	multi-agent	多个智能体分工、评审、协作
人在回路	human-in-the-loop	关键环节保留人工判断与验收
停止条件	stopping condition	明确"任务何时算完成"的规则
显式 / 隐式标识	explicit / implicit marking	可见的 AI 提示 / 文件元数据标识
深度合成	deep synthesis	法规语境下对换脸、合成人声等的统称

附录 B 网页工具速查（2026 年 10 月版）

按"能力"而非"品牌"组织——需要哪种能力，就去对应行找当期可用的工具。所有条目为编写时检索所见，**额度、价格与条款以官方页面为准**；"课堂适用"一列是基于"机房零配置 + 免费可用 + 中文友好"三条标准的建议。

表 B-1　工具速查总表（能力 → 代表工具 → 课堂适用性）

能力	代表工具（以国内为主）	课堂适用性备注
对话 / 写作 / 规划	DeepSeek、豆包、通义千问、Kimi、智谱清言、文心一言、腾讯元宝	网页版即可用，基础功能免费；建议同任务用两家对照（第 1 章"跨模型对照"）
图像生成与编辑	即梦、通义万相、豆包、可图、混元；国际对照：Midjourney、FLUX 系、Adobe Firefly	国产工具免费额度足够课堂练习；编辑功能（局部重绘/扩图）是重点（第 2 章）
视频生成	可灵、即梦、海螺、Vidu、通义万相视频、腾讯混元视频；国际对照：Veo、Sora、Runway、Pika	免费额度多为"每日数次"；课堂策略：小分辨率短片段多试，成片时再升级（第 3 章）
三维生成	腾讯混元 3D、Tripo、Rodin、Meshy、TRELLIS 系开源模型	生成快、拓扑常需修复；导出格式选 glb/obj（第 4 章）；开源模型适合进阶（附录 F）
配音 / TTS	剪映（免费不限量，首选）、讯飞智作、MiniMax 语音、CosyVoice（开源）	"公共音色 + 自己的声音"是课堂安全区；克隆他人声音需授权（第 1 章判例）
音乐 / 音效	海绵音乐、天工 SkyMusic、MiniMax 音乐；国际对照：Suno、Udio	注意条款差异（权属与商用许可各家不同）；课堂用"描述→生成→试用"流程
剪辑 / 后期	剪映、必剪、达芬奇（免费版）	剪映与前面的生成工具衔接最顺；智能字幕、AI 配音内置（第 3 章）
智能体平台	扣子（Coze）、Dify、文心智能体平台、腾讯元器	零代码编排，网页直接搭工作流；注意平台的免费额度与发布渠道（第 8 章）

> **🎯 注册与准备清单（第 1 周作业用）**
>
> ① 1 个对话助手（主力）+ 1 个备选（对照用）；② 1 个图像工具；③ 1 个视频工具（可稍后再开）；④ 剪映（桌面或手机版）；⑤ 云盘/笔记（存"创作档案"四件套：成品 / 参数记录 / 对照实验 / 素材来源）。**注册时用真实手机号，注意未成年人条款与实名要求；所有工具的条款页面建议收藏**——第 8 章你会用到它们。

附录 C 提示词模板库（可直接复制）

十二个经课堂验证的模板，覆盖全流程。用法：把［占位符］替换成你的内容，先跑三版，再按"四维评估法"迭代。**网页版点"复制提示词"按钮即可。**

✎ C-1 脚本 / 解说词创作

你是一位纪录片解说词撰稿人，风格含蓄克制、以景写情。

任务：为 ［作品名 / 主题］ 撰写 ［时长，如 60 秒］ 的解说词。

约束：
- 围绕 ［核心意象 / 原典范围，如"张若虚《春江花月夜》前八句"］，不要展开太远；
- 不堆砌形容词，不出现"震撼""绝美"这类词；
- 总字数 **［150 字］**，以短句为主，便于配音朗读；
- 不虚构历史事实，拿不准的细节标注"待核实"。

输出格式：分为 **［4 个镜头］**，每个镜头两行——
画面：（一行画面描述）
旁白：（一行解说）

参考示例：
画面：夜江全景，一轮明月自潮水中升起
旁白：潮水漫过春天的江面，月亮从水底升了起来。

模式 普通　温度 0.8

✎ C-2 分镜脚本表（在解说词之后使用）

以下是我们已定稿的解说词与风格说明。
【风格说明】［含蓄克制；画面留白多；色调由冷转暖］
【定稿解说词】
［粘贴解说词］

任务：把它拆解为分镜头脚本表。

约束：
- **［8–10］** 个镜头，总时长 ［60］ 秒，单镜头 4–10 秒；
- 每镜头标注：景别（远/全/中/近/特）、机位与运动（固定/推/移/摇）、画面主体；
- 为每镜头写一条可执行的 AI 生成提示词草稿（中文，≤40 字）；
- 标注该镜头的可能风险（如"水面反光一致性难，建议视频生成"）。

输出格式：Markdown 表格，列为：镜头号 | 时长 | 景别 | 机位运动 | 画面主体 | 生成提示词草稿 | 风险备注

模式 深度思考

✎ C-3 文生图（图像生成通用骨架）

［主体：月夜下的江面与孤舟］，［动作/状态：小舟随微波轻晃］，［环境：远山含黛、薄雾］，［构图：16:9 电影感，主体在右三分之一处］，［光线：月光为主光源，冷色调］，［风格：水墨动画质感，笔触柔和］，［细节约束：画面干净，不要文字（no watermark, no text）］

负向：［多余的手指、肢体畸变、文字与水印、过曝、塑料感］

比例 16:9　引导强度 中（5–7）　步数 30–50

✎ C-4 图生图 / 局部重绘（保持主体，改一处）

以参考图为底：**只修改［指定区域 / 指定元素］**，其余保持（人物、构图、光影、色调）不变。
【要改的】［把左侧的树改成垂柳，姿态更柔］
【不能变的】［人物姿态、画面构图、整体色调、笔触风格］
【重绘强度建议】［0.3–0.5，越低越保守］

重绘强度 0.3–0.5　注意 重绘前改变"模型可见信息"

✎ C-5 视频生成（含运镜描述）

［镜头语言：中景，镜头缓慢上摇（tilt up）］，［主体：诗人立于船头，衣袖随风轻扬］，［环境：夜江、明月、薄雾］，［动作强度：轻微而持续］，［风格：水墨动画，冷色调，画面稳定］

约束：时长 ［5 秒］；避免快速动作与镜头切换；保持人物面部稳定。

条件 图生视频（首帧用 C-3 结果）　运动幅度 低–中

✎ C-6 三维生成（文生 3D / 图生 3D）

［物体：中国传统楼台（歇山顶，两层）］，［风格：写实低多边形，纹理简洁］，［用途：游戏场景资产，需要合理面数］，［细节：结构完整，不要虚幻的装饰］

约束：输出 ［glb］ 格式；纹理分辨率 ［2K］；如支持，指定"单一物体、独立放置"（便于后续绑定与摆放）。

检查 多视角截图 + 面数 + 拓扑

✎ C-7 配音脚本标注（给 TTS 的"可朗读文本"）

请把以下解说词改写为"给配音用"的版本：
- 把长句拆成 15 字以内的短句；
- 用标点分级控制停顿：逗号=短停，破折号=中停，句号=长停；
- 为多音字与专有名词标注读音（用拼音）；
- 保持原意与风格，不要添加新内容。

原文：［粘贴解说词］

用途 剪映 / TTS 工具　语速 0.9× 试听

✎ C-8 配乐生成（五要素需求）

用途： [短片开场 60 秒旁白底衬，纯器乐]
情绪走向： [由静谧微凉，到开阔明亮，结尾回到安静]
乐器： [箫或古琴主奏，少量低音弦乐铺底，不用鼓]
时长与结构： [60 秒整，末尾 3 秒自然收束（渐弱不截断）]
禁忌项： [不要人声、不要电子鼓点、不要突兀转调]

生成 3 版候选，人工挑　注意 条款中的权属与商用许可

✎ C-9 场景需求 → 结构化资产清单

背景： [我们要搭一个"宋时江畔夜市"场景，2 分钟短片用，卡通写实混合风格，性能预算中等（总面数希望控制在 30 万以内）]

任务：把需求转成"场景规格 + 资产清单"。

输出格式：
一、场景规格（JSON）：{ "主题": ..., "功能分区": [...], "动线": "...", "光照": "...", "色调": "..." }
二、资产清单（Markdown 表）： 资产名 | 类型（地形/建筑/道具/植被） | 数量 | 生成方式（文生3D/图生3D/天空盒生成/手绘） | 预估面数 | 优先级（必做/加分） | 风险备注

模式 深度思考

✎ C-10 任务分解与验收标准（智能体实训）

目标： [在两周内完成 2 分钟古诗词短片《春江花月夜》的数字资产制作]

任务：
1. 把目标分解为可执行、可验收的子任务（≤12 个）；
2. 每个子任务给出：输入（依赖哪些前置产出）、输出（交付什么文件/素材）、验收标准（怎么算完成，尽量可核查）、负责人角色（策划/美术/技术）；
3. 标出关键路径（哪些任务延误会导致整体延期）；
4. 在最后列出"最可能失败的三个环节"及备选方案。

输出格式：Markdown 表格 + 关键路径说明 + 风险清单。

模式 深度思考

✎ C-11 跨模型对照实验记录

任务：把下面的原始需求分别向两个不同模型提问，记录差异。

原始需求： [粘贴你的一句话需求]

对照记录模板（输出成表格）：
维度 | 模型 A 表现 | 模型 B 表现 | 差异结论
忠实度（是否扣题）
可用性（能否直接用）
表现力（是否有亮点）
可控性（重复生成是否稳定）
一句话结论： [差异主要来自什么？提示词？模型风格？]

温度 两模型设为一致

✎ C-12 技术说明撰写（期末用）

以下是我们作品的技术记录： [粘贴参数记录、对照结果、修改痕迹]

任务：把它整理成"技术说明"，面向不了解我们创作过程的读者。

结构要求：
1. 作品概述（100 字内）；
2. 创作流程（一条流水线：每类资产用了什么工具、什么参数）；
3. 参数对照（至少 2 组 A/B 对照：变量、结果、结论）；
4. 人工修正清单（AI 生成了什么、我们改了什么、为什么）；
5. AI 参与程度说明（哪些环节以 AI 为主、哪些以人为主）；
6. 合规说明（素材来源、音色授权、AI 标识）。

约束：所有数据必须来自记录，不得编造；语气客观，不夸张。

用途 占总评 20% 的交付物

附录 D 期末综合作品指南与验收标准

◆ D-1 作品要求（一句话版）

用至少两类数字媒体，完成一部能公开展示的综合作品（推荐形态：1–3 分钟的短片，或互动型媒体作品），并提交配套的"技术说明与参数对照"。小组提交须由每位成员注明个人贡献。

推荐的"稳过线"配置（融合四类以上媒体，覆盖全部考核点）：**短片 = 文本脚本 + AI 配音 + 分镜图像 + 生成视频镜头 + 配乐 + （可选）三维资产或智能体工作流。**全书贯穿案例《春江花月夜》就是这个配置的完整示范。

◆ D-2　时间线：11 周怎么走

表 D-1　期末作品推进时间线（与教学周同步）

周次	里程碑	交付物（存入创作档案）	对应章节
第 1–2 周	选题立项 + 脚本与配音	主题说明、分镜脚本初稿、解说音频	第 1 章
第 3 周	视觉设定	角色设定图、4–6 张分镜关键帧、风格参考板	第 2 章
第 4 周	镜头成片	4–6 段镜头视频、失败案例分析与改进	第 3 章
第 5 周	三维资产	模型初稿、多视角截图、缺陷记录	第 4 章
第 6 周	角色动画	动画片段、前后对比、修正记录	第 5 章
第 7–8 周	动作迁移	姿态提取结果、迁移视频、错误分析	第 6 章
第 9 周	场景整合	场景初稿、结构化资产清单	第 7 章
第 10 周	工作流与修订	智能体任务分解表、场景修订稿	第 8 章
第 11 周	公开展示与验收	成片、协作记录、个人贡献说明、技术说明	全书

◆ D-3　验收评分表（对照课程总成绩构成）

期末 70% 的三个部分逐项展开如下。**建议把它当作"提交前自查清单"逐条打勾**：

表 D-2　综合作品评分要点（40%）

维度	优秀（9–10）	合格（6–8）	不达标（<6）
完成度	作品完整、时长达标、两类以上媒体融合自然	基本完整，媒体融合较生硬	缺项明显或未完成
技术与方法的运用	参数对照充分、会根据评估迭代方法	用了课程方法，但对照实验较少	只有"抽卡式"生成，无方法痕迹
审美与表达	有明确的风格与表达意图，人工修正到位	完成度尚可，风格不统一	素材堆砌，无整体感
合规与诚信	素材来源、授权、AI 标识齐全，贡献说明清晰	基本合规，个别记录缺失	来源不清或未标识

表 D-3　技术说明与参数对照评分要点（20%）

维度	优秀（9–10）	合格（6–8）	不达标（<6）
流程还原度	读者能照着复现主要环节（工具、参数、顺序齐全）	流程大致清楚，个别参数缺失	只有结果，没有过程
对照实验质量	≥2 组规范对照（单变量、有结论、能解释原因）	有对照但变量不单一或结论空泛	没有对照实验
人工修正说明	清楚说明"AI 产出 vs 人工修改"及修改理由	提到了修改，理由简略	未区分 AI 与人工

表 D-4　公开展示与个人贡献评分要点（10%）

维度	优秀（9–10）	合格（6–8）	不达标（<6）
展示表达	讲清"做了什么、为什么这样做、学到了什么"	讲清了作品，但技术逻辑含糊	照稿念或超时/严重不足
互评参与	对他人作品的评价具体、有建设性	参与互评但意见笼统	缺席或敷衍
个人贡献	贡献说明具体到任务与产出，可核查	说明笼统但可对应	无法区分个人贡献（挂名嫌疑）

⚠️ **期末高频失分点（往年经验）**

① **记录缺失**——"技术说明"写不出来的根本原因，往往是平时没有存参数；② **对照实验不规范**——同时改了两个变量，结论不成立；③ **合规硬伤**——素材来源不清、AI 内容未标识（这类问题在"合规与诚信"维度是"一票降档"）；④ **贡献说不清**——小组作品最常见的不及格原因。四条全部可以在平时避免，不必留到期末。

附录 E 合规与授权操作清单

◆ E-1 发布前"三查"清单（打印张贴版）

表 E-1 发布前检查清单

检查项	具体动作	通过标准
一查素材来源	逐项核对素材来源档案：来源 URL、授权方式（自有/公共领域/授权库/AI 生成）、日期	每个素材都有完整来源行；来路不明的一律替换
二查真人权利	排查作品中出现的真人肖像、声音、身体影像；核对授权文件的范围是否覆盖本次发布	有授权或已替换/删除；声音克隆使用合法来源（自己/公共音色/书面授权）
三查 AI 标识	为 AI 生成/合成的图、音、视频添加显式标识（平台标签或画面标注），保留隐式标识（不清理元数据）	标识齐全且符合最低技术要求；发布文案中如实说明 AI 参与

🧪 **关于 AI 标识的"最低技术要求"（来自强制性国标）**

《人工智能生成合成内容标识办法》与配套强制性国家标准（2025 年 9 月 1 日同步施行）对显式标识给出了量化要求，教材转述两个最容易用到的要点：**① 标识文字应包含"人工智能"或"AI"要素，以及"生成"或"合成"要素**（例如"AI 生成""人工智能合成"）；**② 图片的显式标识一般放在边或角，文字高度不低于画面最短边长度的 5%**——就是说，加标识不是"随手写个小字"，而要保证"显著、可被注意到"。视频与音频的标识位置要求见标准原文；各平台（抖音、快手、小红书、B 站等）也陆续上线了"AI 生成内容"声明标签，发布时优先使用平台的声明功能，再叠加画面/文案标识。

◆ E-2 素材来源档案模板

表 E-2 素材来源档案（示例两行，复制到你的创作档案）

素材	类型	来源	授权方式	获取日期	备注
江面镜头（片段 3）	AI 生成视频	即梦（账号：xxx）	平台生成内容（按平台条款）	2026-10-20	已加 AI 标识；参数见参数记录表第 12 条
古琴音色采样（第 8 小节）	AI 生成音乐	海绵音乐	平台生成内容（条款已截图存档）	2026-10-21	未使用真人演奏采样

◆ E-3 授权书模板（声音/肖像使用）

📄 **声音或肖像使用授权书（简版模板）**

兹授权 ［被授权人/小组名称］ 于 ［使用目的，如"《春江花月夜》课程短片制作与课堂展示"］ 中，使用本人 ［声音录音 / 肖像照片 / 视频影像］，用于以下范围： ［创作、修改（含 AI 处理）、课堂展示、课程平台留存］。使用期限： ［自签署日起至课程结束/具体日期］。超出上述范围的使用需另行取得本人同意。

授权人（签名）：____ 联系方式：____ 日期：____

注意：①"用 AI 处理"要单独写明——"录音授权"不等于"声音 AI 化授权"（第 1 章判例）；② 涉及未成年人的，须监护人共同签署；③ 扫描件与原件同效，存入素材档案。

◆ E-4　法规速查（创作者视角）

表 E-3　法规速查表（截至 2026 年 10 月）

法规	施行时间	创作者要记住的一句话
《网络安全法》（修订）	2026-01-01	AI 首次入法：发展与安全并重
《互联网信息服务深度合成管理规定》	2023-01-10	换脸、合成人声、智能写作，必须显著标识
《生成式人工智能服务管理暂行办法》	2023-08-15	使用的工具应依法备案；不得生成违法有害内容
《人工智能生成合成内容标识办法》＋ GB 45438—2025	2025-09-01	发 AI 内容要"亮明身份"，标识有量化要求
《学位法》	2025 年	AI 代写学位论文 = 学术不端
《人工智能拟人化互动服务管理暂行办法》	2026-07-15	AI 角色"陪聊"类服务有专门监管
最高法涉 AI 纠纷案件意见（法发〔2026〕10 号）	2026-09-07	声音权益、标识、数据边界的裁判口径

附录 F　选学：最小 API 调用与本地工作流

本附录是"进阶通道"，**不计入课程考核**：给想走得更远的同学两条路——"用代码调用模型"（最小 API）与"在本机跑模型"（本地工作流）。它们对应课程能力目标里的"最小 API 调用"，也是第 8 章"工具调用"的物理基础。

◆ F-1　API 是什么：一个"点菜窗口"

你平时用的网页版工具，是"餐厅"——进去、坐下、点菜、吃饭，全在餐厅里完成。而 **API**（应用程序接口）是"外卖窗口"：你不是去餐厅，而是把你的需求（菜名、数量、口味）写成一张标准订单递进窗口，厨房做好后把餐盒（数据）递出来。订单的格式叫"请求"（request），餐盒叫"响应"（response）。

为什么创作者要懂这一层？因为**网页版是"一次一道菜"，API 是"批量下单"**：当你要为 20 个镜头各生成 3 版配图、或让程序自动检查素材版权时，手动点 60 次不如写一个脚本。更重要的是，第 8 章的智能体"调用工具"，走的就是 API 这条通道——理解了外卖窗口，你就理解了智能体为什么能"办事"。

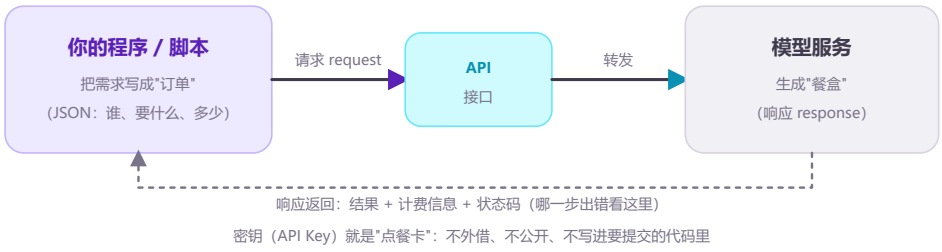


图 F-1　API 的"外卖窗口"模型。请求-响应是全部网络服务的基本形态；第 8 章的"工具调用"正是智能体替你写好订单、递进窗口、再读回餐盒。

◆ F-2　最小调用示例（读懂即可，不要求运行）

下面是一个"最小 API 调用"的样子（以对话类接口为例，各家格式大同小异）。**读它有四个要点**：① 密钥放在请求头里；② 内容用 JSON 描述（模型名、消息列表、参数）；③ 参数就是你第 1 章学过的温度、长度上限；④ 返回值里取出生成的文本。

✎ F-2 • 最小调用（Python 风格伪代码）

```
import requests

resp = requests.post(
    "https://<服务商地址>/v1/chat/completions",
    headers={"Authorization": "Bearer <你的 API Key>"},
    json={
        "model": "<模型名>",
        "messages": [
            {"role": "system", "content": "你是一位纪录片解说词撰稿人"},
            {"role": "user",   "content": "为《春江花月夜》写 60 秒解说词，分 4 个镜头"},
        ],
        "temperature": 0.8,
        "max_tokens": 800,
    },
)

print(resp.json()["choices"][0]["message"]["content"])
```

要点 密钥保密 · 先小后大 · 出错看状态码

⚠️ **三条安全红线**

① **密钥绝不进代码仓库、截图与提交材料**——泄露的密钥会被人盗刷（平台按用量计费）；② 先用最小请求测试，再批量调用——避免"一觉醒来账单爆炸"；③ 调用即代表你同意服务条款：注意数据使用条款与内容政策，涉及未公开素材时优先考虑本地方案（F-3）。

◆ F-3　本地工作流：在自己机器上跑模型

"本地工作流"指在自己的电脑（通常是带独立显卡的机器）上运行开源模型，而不是调用云端服务。它的三个价值：**数据不出本机**（隐私敏感素材的最佳选择）、**不按次付费**（电费换算力）、**完全可控**（可改装、可组合工作流）。代价：需要一定硬件（近年的独立显卡即可起步）、需要动手配置、生成速度取决于显卡性能。

典型的本地玩法有两种：**① 命令行/脚本调用本地模型**——下载开源模型权重（如开放权重的中文模型与语音模型），用推理框架在本机跑，再写脚本批量处理；**② 可视化工作流工具**——用节点式界面（如社区流行的 ComfyUI 生态）把"文生图→放大→重绘→导出"连成一条流水线，鼠标拖拽即可复现复杂流程，也可以保存为模板分享给同学。

📌 **给"想试试"的同学：一条安全的入门路径**

① 先用实验室/自备的独显机器装一个可视化工作流工具，跑通"文生图"最简流程；② 再加载一个开源中文对话模型，感受"本地大模型"的响应速度与能力档位；③ 把第 2 章的"参数对照实验"在本地重做一遍——你会第一次拥有"完全可复现"的种子与参数环境；④ 学有余力再看语音模型（附录 B 中的开源 TTS）。**全程遵守开源许可协议（有些模型允许商用、有些仅限研究，下载页会写明）。**

把六份附录压缩成六句话

1. 术语表按章分组：每章学完"过一遍"，能用自己的话解释就是掌握。
2. 工具速查按"能力"组织：先认能力，再找当期可用的工具——工具会换，能力不会。
3. 提示词模板库是起点不是终点：替换占位符、跑三版、按四维评估迭代。
4. 期末评分表就是验收清单：作品 40%、技术说明 20%、展示与贡献 10%，逐条自查看得见的分。
5. 合规"三查"（素材、真人、标识）是发布前的固定动作；授权要写明"AI 处理"范围。
6. API 是"批量创作"与"智能体办事"的通道；本地工作流是隐私与可控性的后手——选学，但值得知道它存在。

自测（附录综合）

❓ 自测 A1：为什么工具速查表要按"能力"而不是按"品牌"组织？

答案：因为品牌会消失、会改名、会停服（你翻一下 2020 年推荐的"AI 工具清单"就明白了），而"文本转语音""局部重绘""三维生成"这些能力类别，在未来数年都会存在。按能力组织，意味着每次换工具时，你的知识不贬值——你需要做的只是"在对应能力格里换一个工具"，方法（参数语义、评估维度、工作流）全部保留。这也是本书"讲原理、不迷信型号"原则在附录里的体现。

❓ 自测 A2：小组作品里，小舟负责"用 AI 生成了全部视频镜头"，个人贡献说明里这样写行吗？为什么？

答案：不合格。这份说明没有区分"AI 做了什么"与"人做了什么"——按课程要求，个人贡献应具体到**人的判断与劳动**：提示词与参数的确定、对生成结果的筛选（从几版库存里选了哪一版、为什么）、失败案例的处理、与组内其他环节的对接。写"用 AI 生成了全部镜头"等于说"我按了按钮"，既无法核查，也违背"如实标注 AI 参与程度"的要求（对应评分表"个人贡献"维度的"一票降档"风险）。正确写法示例："负责镜头 1–4 的生成：设计并迭代了 3 版提示词；对运动幅度做了低/中两档对照（结果见对照表 7）；从每镜头 6 版候选中依据'时间一致性'筛选；对镜头 3 的闪烁问题用重绘修复 2 次。"

❓ 自测 A3：为什么"录音授权"不等于"声音 AI 化授权"？这对你的创作实践意味着什么？

答案：这是司法明确确立的规则（第 1 章两起判决）：某公司拿到了配音师录音制品的著作权，仍被判侵权——因为"对录音制品的授权并不意味着对声音 AI 化的授权"。声音是人格权益的载体（民法典参照肖像权保护），把它用作 AI 训练的语料、或用来克隆音色，是**另一项独立的用途**，需要单独的授权。实践含义：① 自己写授权书时，把"A I 处理/合成/克隆"单列为授权范围；② 作为使用者，即使素材"看起来来源合法"，也要确认授权范围是否覆盖 AI 化使用；③ 课堂安全区永远是"自己的声音 / 平台公共音色库"。

F-4　本书插图的生成记录（可复现）

按本书自己教的规矩，把教材插图的生成信息如实记录如下，供同学们复现与研究：

表 F-1　本书 AI 生成插图的生成记录

项目	内容
工具	本地部署的 ComfyUI + Qwen-Image 2.1（开源权重，本地推理）
生成方式	文生图（text-to-image）；未使用任何他人作品的参考图或图生图
关键参数	采样步数 25（对照实验组除外，见下）；CFG 1.0（该模型为蒸馏/低引导工作点）；分辨率 1–2 MP（按用途）
对照实验组	第 2 章的 图 2-8 与 图 2-9 为**受控对照**：分别只改变提示词写法（同种子 20261013）与采样步数（同种子 20261012），其余参数完全一致
统一画面基调	「电影感构图 · 青蓝色调 · 克制含蓄 · 大量留白 · 东方美学」，并要求画面不出现任何文字与水印
使用范围	仅作为教材内部的教学示例；素材为原创生成，不涉及第三方 IP 形象
标识情况	全部插图在图注中标注「AI 生成」；未清理生成文件的元数据

🎯 为什么要写这一节

因为这是本书反复强调的三件事的示范：① **记录**（参数与种子可复现——你能用同样的记录重做出类似的图）；② **标识**（AI 生成就写明 AI 生成，不冒充摄影或手绘）；③ **边界**（不模仿第三方 IP、不使用他人素材）。 你们期末作业的技术说明，写的就是这份东西。

◆ 拓展阅读

- 你最常用的那个工具"官方文档"的"服务条款"与"内容政策"页——本课程要求至少精读一遍你主力工具的这两页（多数同学从没点开过）。
- 《人工智能生成合成内容标识办法》与 GB 45438—2025 标准文本——发布作品前对照操作。
- 开源社区（如模型托管平台的开源模型页面）——看看"模型卡片"（model card）里都写了什么：能力、限制、许可、伦理声明，这是技术透明的最好样本。
- 本课程超星学习通的教师自编资料：《文生图扩散模型解析》《文本到三维模型进展分析》《视频生成技术演进》——与各章配套的延伸阅读。