文生视频大模型：评测科学、长视频、世界模型、后训练与安全（下册）

从“会生成”到“可验证、可交互、可对齐、可治理”

面向具备深度学习基础的计算机博士生

2026-07-12

目录

下册前言：生成系统的终点不是样片，而是可信证据

上册建立了文生视频的概率建模、Video VAE、扩散/流、Video DiT、MLLM Planner 与基础指标；中册进一步解剖了 Wan、Bernini、数据工程、训练系统、推理成本与复现实验。下册处理研究中更困难、也更容易被“漂亮 demo”掩盖的问题：

- 如何判断一个指标真的测到了目标能力，而不是模型学会了取巧？
- 为什么 FVD、CLIP 相似度或单一 VLM judge 不能独立支撑“世界模型”结论？
- 如何把动作顺序、状态变化、物理规律和因果关系转成可检验的结构？
- 为什么短视频效果可以很好，而长视频仍会身份漂移、事件遗忘和因果崩溃？
- 文生视频模型与动作条件世界模型究竟差在哪里？
- 怎样训练视频奖励模型，并用 DPO、DenseDPO、GRPO 或 reward gradient 做后训练？
- 为什么 reward 提高后，人类偏好、物理真实性或多样性反而可能下降？
- 面向深度伪造、记忆泄漏、版权、隐私与错误来源归因，研究者应建立什么防线？
- 一个计算资源有限的博士生，怎样把这些问题收敛为可证伪、可复现、可发表的课题？

本册的中心思想是：**视频生成是一项测量科学、序列决策问题和社会技术系统问题，而不只是采样若干视觉结果**。一个可信结论至少需要四类证据：

$$\text{能力定义} + \text{受控实验} + \text{统计不确定性} + \text{失败与风险审计}.$$

资料时点：论文、技术报告、开源基准与政策资料核对至 2026 年 7 月 12 日。2026 年的新论文大多仍是预印本，文中会区分“论文公开报告”“由公式/代码推导”“研究性判断”。排行榜、商业模型版本和法规实施细节会继续变化，复现实验应锁定版本、日期与评测脚本哈希。

本册学习目标

完成下册后，读者应能够：

1. 将视频生成评测视为潜变量测量问题，分析 construct validity、reliability、calibration 与 statistical power；
2. 设计包含多 seed、配对比较、层级 bootstrap 和多重比较修正的人类评测；
3. 正确解释 FVD、JEDi、CLIP/VLM judge、VBench、VBench-2.0、VideoPhy-2 等工具的适用边界；
4. 将复杂 prompt 拆解为实体、属性、关系、事件和时间边，并建立过程感知评分；
5. 推导长视频的分块自回归、重叠窗口、记忆更新、self-forcing 与长程误差累积；
6. 区分开放环文生视频、动作条件预测、可交互世界模型和用于规划的世界模型；
7. 建立 pointwise、pairwise、dense/local 和 process-level 视频奖励模型；
8. 推导视频 DPO、DenseDPO、LocalDPO、reward gradient 与 GRPO 的统一目标；
9. 识别 reward hacking、likelihood displacement、mode collapse 与安全对齐回退；
10. 建立涵盖数据、人物授权、输入输出过滤、C2PA/水印、红队和事件响应的治理体系；
11. 将上述内容落实为 90 天研究计划、实验矩阵、代码工具和论文证据链。

新增统一符号

沿用上、中册的 $\mathbf{x}, \mathbf{z}, \tau, v_\theta, E_\phi, D_\psi, N, K$ 等符号。本册新增：

符号	含义
$q \sim \mathcal{Q}$	评测 prompt 或测试任务，$\mathcal{Q}$ 为目标任务分布
$m \in \{1, \dots, M\}$	待比较的视频生成模型
$s \in \{1, \dots, S\}$	随机种子或独立生成样本
$r \in \{1, \dots, R\}$	人类评审或自动评测器
Y_{qsmr}	在 prompt、seed、模型、评审组合上的观测分数
$\xi_m(q)$	模型 m 在任务 q 上的潜在真实能力
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	视频事件图，节点为实体/状态/事件，边为时间或因果关系
$\mathbf{s}_t$	世界模型或长视频系统在时刻 t 的隐状态
$\mathbf{a}_t$	用户、代理或机器人动作
$\mathbf{o}_t$	可观测视频/传感器输出
$\mathbf{M}_k$	第 k 个视频 chunk 后的长期记忆
$R(\mathbf{x}, q)$	视频奖励函数，可能为多维向量
π_θ	后训练中的生成策略/模型
π_{ref}	冻结参考模型

符号	含义
β	DPO/偏好优化温度或 KL 强度；上下文会明确
A_i	GRPO 中第 i 个候选的组内标准化优势
δ_{phys}	物理规律残差或违反程度
α	显著性水平；不与注意力权重混用

三类陈述的标记原则

证据等级

A 类：公开证据。论文、官方技术报告、官方规范或开源实现直接披露。**B 类：可复算推导**。由公开结构、公式或配置推导，如 token 数、统计置信区间、成本估算。**C 类：研究性判断**。机制解释、课题建议或未来趋势，需要实验验证，不能当作已证实事实。

第 28 章　把视频评测当作测量科学

视频生成论文常把评测写成“运行若干指标并比较均值”。但从科学方法看，我们真正想测的是不可直接观察的能力，例如：

- 文本条件遵循；
- 时序一致性；
- 组合泛化；
- 物理正确性；
- 长程叙事一致性；
- 可控性；
- 人类偏好和任务效用；
- 安全性与公平性。

这些都是**潜在构念（latent construct）**。FVD、CLIPScore、VLM judge、人类二选一只是这些构念的带噪代理。

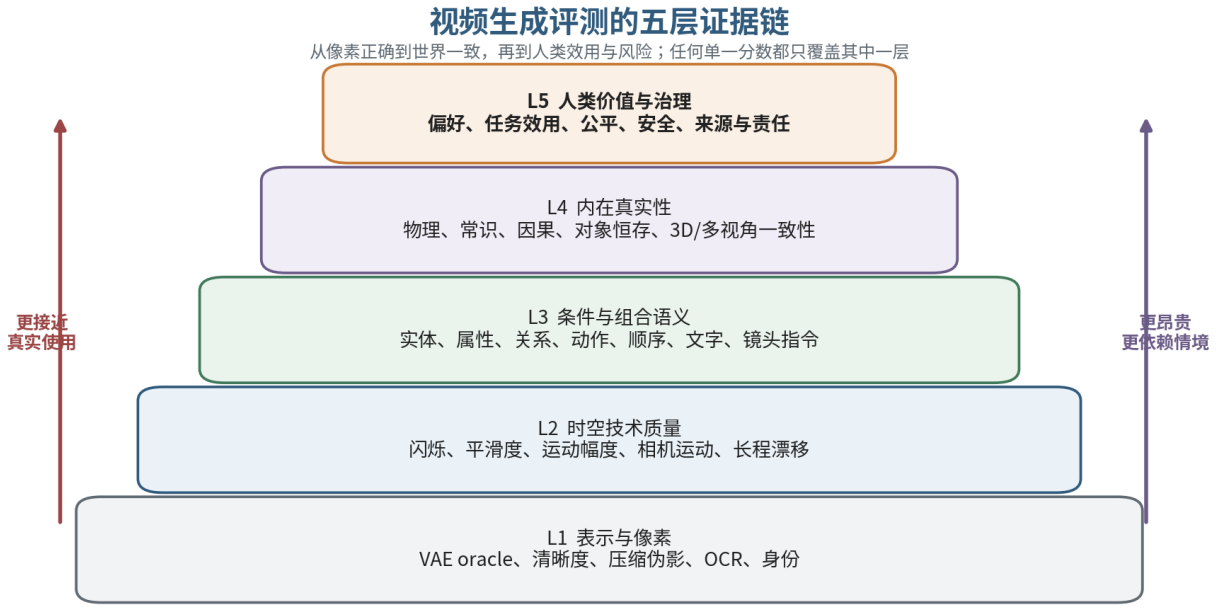

图 1: 视频生成评测的五层证据链

28.1 从“指标”回到“目标构念”

设模型 m 在 prompt q 上存在潜在能力 $\xi_m(q)$，评测器 j 输出：

$$Y_{m,q,j} = f_j(\xi_m(q), \mathbf{c}_q) + b_j(\mathbf{c}_q) + \varepsilon_{m,q,j},$$

其中：

- f_j 是评测器对目标能力的响应函数；
- $\mathbf{c}_q$ 是 prompt 类型、视频长度、人物数量、相机运动等协变量；
- b_j 是系统偏差；
- ε 是随机误差。

理想指标应近似单调：能力更高时得分不应系统性降低。但实际中 f_j 可能只在局部单调。例如 CLIP 相似度对“猫在桌上”有效，对“猫先跳下桌，再绕到椅子后面”则可能主要响应第一帧主体。

核心结论

评测设计的第一步不是选择指标，而是写出一个可证伪的能力定义：**输入是什么、期望行为是什么、容许变化是什么、失败条件是什么、哪个观察单位支持该判断**。

28.2 五类测量质量

28.2.1 构念效度（construct validity）

指标是否真的测到了目标能力？可进一步分为：

- **内容效度**：测试集是否覆盖目标能力的主要子维度；
- **收敛效度**：不同测量同一能力的指标是否一致；
- **区分效度**：指标是否与无关能力分离；
- **已知组效度**：对人为构造的明显好/坏样本，指标能否正确排序。

例如“运动质量”至少要区分：主体运动、相机运动、背景动态、闪烁、运动幅度、运动合理性。只用平均光流会把相机抖动误判为主体活跃。

28.2.2 可靠性（reliability）

同一对象重复测量能否得到相近结果？

- test-retest：相同视频重复评分；
- inter-rater：不同评审的一致性；
- inter-seed：相同 prompt 不同生成 seed 的方差；
- implementation reliability：不同解码器、视频预处理、特征抽取版本是否一致。

对连续评分，可使用组内相关系数（ICC）；对类别标签，可使用 Cohen/Fleiss κ 或 Krippendorff α。但“一致”不代表“正确”：所有评审都受同一偏差影响时，可靠性高而效度低。

28.2.3 灵敏度与分辨率

若模型改进很小，指标是否有足够 power 检测？定义标准化效应量：

$$d = \frac{\bar{Y}_A - \bar{Y}_B}{s_{\text{pooled}}}.$$

当生成方差很高而 prompt 数很少时，论文中的 0.5 分提升可能完全落在随机波动内。对视频生成，增加 **prompt 数**通常比在极少 prompt 上增加大量 seed 更能提高外部效度；但每个 prompt 至少需要多个 seed 才能估计生成随机性。

28.2.4 校准（calibration）

自动 judge 输出 0.9 是否意味着约 90% 的人类认为正确？二分类评测器可用 Expected Calibration Error：

$$\text{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n} \left|\text{acc}(I_b) - \text{conf}(I_b)\right|.$$

VLM-as-a-Judge 经常出现高置信错误，特别是在快速动作、遮挡、否定关系、镜头切换和文字细节上。因此应在目标视频域上做 calibration，而不是直接使用模型的自然语言自信度。

28.2.5 鲁棒性与可迁移性

评测结论是否对以下扰动稳定：

- 编码格式、压缩质量和抽帧率；
- 视频分辨率与宽高比；
- prompt 改写、同义词、语序；
- judge 模型版本；
- 不同文化、场景和人物群体；
- 开放源模型与闭源模型的输出风格。

若一个指标仅在特定分辨率或特定 VLM 上有效，论文应将其称为“该协议下的代理得分”，而不是普遍能力。

28.3 评测的观测单位与随机变量

完整观测可写为：

$$Y_{q,s,m,r} = \mu + \alpha_m + u_q + v_r + (\alpha u)_{m,q} + \epsilon_{q,s,m,r},$$

其中：

- α_m：模型固定效应；
- u_q：prompt 难度随机效应；
- v_r：评审严格度随机效应；
- $(\alpha u)_{m,q}$：模型与 prompt 类型的交互；
- ϵ：seed 与观测噪声。

最常见的统计错误是把同一 prompt 下的多个 seed 当作完全独立样本，从而夸大有效样本量。若目标是推广到新 prompt，推断单位应优先是 prompt 或 prompt family。

28.3.1 方差分解

定义总方差：

$$\mathrm{Var}(Y) = \sigma_q^2 + \sigma_r^2 + \sigma_{m\times q}^2 + \sigma_s^2 + \sigma_\epsilon^2.$$

通过 pilot study 估计各项，可以回答：

- 是否应增加 prompt 数?
- 是否应增加每 prompt 的 seed?
- 是否需要更多评审?
- 哪类 prompt 造成模型排序不稳定?

若 σ_q^2 很大，固定 100 个简单 prompt 的排行榜并不代表真实用户分布；若 σ_s^2 很大，只展示最佳 seed 会严重乐观。

28.4 Prompt 集不是问题清单，而是抽样设计

评测目标是估计：

$$\mathcal{S}_m = \mathbb{E}_{q\sim\mathcal{Q},\, s\sim\mathcal{S}}\left[g(\mathbf{x}_{m,q,s}, q)\right].$$

因此 prompt 集应近似目标分布 $\mathcal{Q}$，或显式采用分层权重：

$$\hat{\mathcal{S}}_m = \sum_{k=1}^{K} w_k \frac{1}{|\mathcal{Q}_k|S} \sum_{q\in\mathcal{Q}_k}\sum_{s=1}^{S} Y_{q,s,m}.$$

推荐至少按以下维度分层：

维度	例子
实体数	单主体、双主体、多人/多物体
动作复杂度	静态、单动作、组合动作、交互动作
时间结构	同时、先后、条件触发、循环、持续状态
空间关系	左右、前后、包含、遮挡、远近变化
物理类型	刚体、液体、布料、烟火、破碎、碰撞
相机	固定、平移、推拉、环绕、手持、切镜
文本	标牌、字幕、物体表面文字、动态变化
人物	身份、服装、手部、群体互动、文化情境
时长	2–4 秒、5–10 秒、30 秒、多镜头分钟级

28.4.1 正例、最小对照和反事实对照

对复杂语义，单一 prompt 不足以确认模型是否理解。应构造三元组：

1. 正例：`红球先滚过蓝球，随后蓝球开始移动`；
2. 顺序反转：`蓝球先移动，随后红球滚过它`；
3. 属性交换：`蓝球先滚过红球...`。

若指标对三者给出近似分数，它大概率只检测“红球、蓝球、运动”是否出现，而未检测关系与顺序。

28.4.2 Benchmark leakage

公开 prompt 长期被用于模型调参与展示后，会成为训练分布的一部分。缓解方法包括：

- 隐藏一部分模板和实体组合；
- 每次评测程序化生成等价新 prompt；
- 使用语义结构而非固定句子定义测试；
- 维护时间戳和模型发布前后的 holdout；
- 对生成结果做训练视频近邻检索；
- 同时报告 public-dev 与 private-test。

28.5 指标的典型失效图谱

图 2: 指标的典型失效与诊断补丁

28.5.1 Goodhart 定律的评测版本

当代理指标成为训练目标后，

$$\arg\max_{\theta} \mathbb{E}[R_{\text{proxy}}] \nRightarrow \arg\max_{\theta} \mathbb{E}[U_{\text{human}}].$$

模型可能通过以下方式提高分数：

- 将主体放大并居中以提高图文相似度；
- 减少运动以提高时间一致性；
- 用镜头运动制造高光流；
- 重复审美模板以提高美学 reward；
- 将复杂事件压缩为一个静态关键帧；
- 生成评测器熟悉的视觉风格；
- 牺牲少数群体或域外场景来提高平均分。

因此任何可训练的自动 reward 都必须配置 held-out evaluator、人评和失败样本审计。

28.6 一份论文级评测协议

一个可复现协议至少公开：

```yaml
benchmark:
  name: private_event_graph_v1
  prompt_manifest_sha256: ...
  strata: [entity_count, temporal_relation, physics, camera, duration]
  prompts_per_stratum: 50
  seeds_per_prompt: 4

generation:
  checkpoint_sha256: ...
  sampler: euler
  steps: 30
  cfg: 5.0
  frames: 81
  fps: 16
  size: [832, 480]
  prompt_rewrite: disabled
  negative_prompt: frozen_v1

automatic_eval:
  preprocessing_commit: ...
  feature_models: [videomae_v2, clip_l14, custom_event_vlm]
  judge_temperature: 0
  repeated_judgment: 3
  calibration_set_sha256: ...

human_eval:
  design: paired_blind
  randomize_left_right: true
  raters_per_pair: 5
  attention_checks: true
  analysis_unit: prompt
  statistic: bradley_terry_plus_prompt_bootstrap
  bootstrap_replicates: 10000
```

常见误区

“我们使用官方设置”不是充分描述。官方仓库、模型、依赖和评测器会更新；必须保存 commit、checkpoint hash、视频预处理参数和 prompt manifest。

28.7 本章自检

练习与自检

1. 一个模型把平均光流提高 402. 100 个 prompt，每个 8 个 seed，为什么统计样本量通常不能直接写成 800? 3. 为“物体恒存”设计正例、反事实对照和可定位失败帧的评分。4. 比较 reliability 与 validity：哪一个可以很高而另一个很低？给出视频评测例子。5. 写出你研究任务的目标分布 Q，并说明当前公开 benchmark 与它的偏差。

第 29 章 分布指标、语义指标与人类偏好的高级用法

上册给出了 FID/FVD、IS、CLIP 类指标的定义。本章不重复基础推导，而关注它们在现代视频生成中的统计性质、适用边界与组合方式。

29.1 Inception Score：为什么几乎不能独立评价开放域视频

对分类器后验 $p(k \mid \mathbf{x})$，IS 为：

$$\mathrm{IS} = \exp\left(\mathbb{E}_{\mathbf{x}} D_{\mathrm{KL}}(p(k \mid \mathbf{x}) \| p(k))\right).$$

它同时偏好：

- 单样本分类置信度高；
- 整体类别分布多样。

但它不使用真实视频分布，也不直接测文本条件、动作或时间一致性。分类器类别之外的生成质量不会得到正确响应。现代开放域 T2V 中，IS 更适合作为历史兼容指标，而不是主要结论。

29.2 FVD：高斯近似、有限样本和特征域偏差

令真实视频特征与生成视频特征分别近似：

$$\phi(\mathbf{x}_r) \sim \mathcal{N}(\mu_r, \Sigma_r), \qquad \phi(\mathbf{x}_g) \sim \mathcal{N}(\mu_g, \Sigma_g).$$

FVD 使用两高斯间的 Fréchet 距离：

$$\mathrm{FVD} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2})^{1/2}\right).$$

29.2.1 它实际测了什么

FVD 由特征抽取器决定。若 ϕ 更关注动作类别，则 FVD 对内容和动作类别敏感；若对精细纹理、文字或真实物理不敏感，FVD 也不会补足。近年的分析指出，经典 FVD 可能存在显著内容偏置；新的工作尝试使用更适合视频表示的特征、联合 embedding 或流式统计替代。

29.2.2 有限样本偏差

均值与协方差的估计在小样本下有偏，且矩阵平方根非线性放大误差：

$$\mathbb{E}[\widehat{\mathrm{FVD}}_n] \neq \mathrm{FVD}(P_r, P_g).$$

因此比较时必须固定：

- 样本数量；
- 视频长度和抽帧；
- resize/crop；
- 颜色空间和编码；
- 特征模型版本；
- 真实参考集；
- 随机 seed；
- 数值实现。

推荐绘制 FVD 随样本数 n 的曲线，或对 prompt/sample 重采样得到 CI，而不是只给一个数。

29.2.3 条件 FVD 与分层 FVD

开放域总体 FVD 可能被类别分布掩盖。可按 prompt 类型 k 分层：

$$\mathrm{FVD}_{\mathrm{strat}} = \sum_k w_k \,\mathrm{FVD}(P_r^k, P_g^k).$$

但每层样本过少时协方差不稳定。实践中可使用 shrinkage covariance、低维 PCA、MMD/KVD 或逐样本语义指标补充。

29.3 Kernel Video Distance、MMD 与 JEDi 类方法

最大均值差异：

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{x,x'\sim P}k(x,x') + \mathbb{E}_{y,y'\sim Q}k(y,y') - 2\mathbb{E}_{x\sim P,y\sim Q}k(x,y).$$

优点是无需高斯假设，可使用多项式或 RBF kernel。无偏 U-statistic 估计为：

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)}\sum_{i\neq j}k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{i\neq j}k(y_i,y_j) - \frac{2}{nm}\sum_{i,j}k(x_i,y_j).$$

JEDi 等方向强调使用更现代的视频表征与更稳健的分布比较。关键仍是：**更换距离公式不能自动修复错误特征**。应同时验证特征对人工扰动的单调响应，例如：

- 帧打乱；
- 局部闪烁；
- 主体替换；
- 运动冻结；
- 速度反转；
- 物体穿透；
- 文本内容保持但动作顺序错误。

29.4 CLIP-based 对齐：从整段相似度到原子命题

最简单的帧平均 CLIP：

$$S_{\mathrm{CLIP}} = \frac{1}{T}\sum_{t=1}^{T}\frac{\langle f_I(\mathbf{x}_t), f_T(q)\rangle}{\|f_I(\mathbf{x}_t)\|\|f_T(q)\|}.$$

它适合主体和静态属性，但不可靠地处理：

- “先…再…”；
- 否定；
- 计数；
- 细粒度左右/前后；
- 动作施事与受事；
- 跨镜头身份；
- 物体状态变化。

29.4.1 命题分解

将 prompt 转为原子命题集合：

$$\mathcal{P}(q) = \{p_1, p_2, \dots, p_J\},$$

例如：

```
q: 一个穿红衣的人把蓝色杯子递给穿白衣的人，然后后者把杯子放到桌上。

p1: 存在红衣人 A
p2: 存在白衣人 B
p3: 存在蓝色杯子 C
p4: A 将 C 递给 B
p5: p4 发生在 B 把 C 放到桌上之前
p6: C 在动作之间保持同一身份与颜色
```

评分不应简单平均；必要条件可使用 conjunction：

$$S_{\text{all}} = \prod_{j \in \mathcal{J}_{\text{required}}} \mathbb{I}[p_j \text{ satisfied}].$$

为减轻单个 judge 错误，可输出逐命题证据帧、置信度和不确定标签，而不是要求 VLM 一次给 0–100 总分。

29.5 VLM/MLLM-as-a-Judge：结构化比自由评分可靠

一个两阶段 judge 可写为：

$$\hat{c} = \text{VLM}(\mathbf{x} \mid S_v), \qquad \hat{y} = \text{LLM}(\hat{c}, q \mid S_l).$$

或者直接使用多问题 VQA：

$$S(\mathbf{x}, q) = \frac{1}{J} \sum_{j=1}^{J} \text{VQA}(\mathbf{x}, Q_j).$$

更可靠的实践：

1. 每个问题只测一个原子事实；
2. 允许 `unknown/not visible`，避免强迫猜测；
3. 要求证据时间段；
4. 同时问正向与反向问题；
5. 固定 temperature；
6. 多次判断并测 self-consistency；
7. 在人工标注集上测 AUC、F1、ECE 与 subgroup gap；
8. judge 与生成模型家族尽量独立；
9. 记录 judge 版本和 prompt；
10. 不把 judge 的解释文本当作真实视觉证据。

29.6 VBench 系列与通用基准

29.6.1 VBench / VBench++

VBench 将视频质量拆成多个维度，而不是给单一总分；VBench++ 扩展到文生视频、图生视频和可信性相关维度。它们适合做广覆盖 profile，但总分依赖维度归一化和权重，不能替代任务特定诊断。

29.6.2 VBench-2.0

VBench-2.0 将重点从“表面真实”推进到“内在真实性”，覆盖 Human Fidelity、Controllability、Creativity、Physics、Commonsense，并进一步拆分细粒度能力。其方法结合通用 VLM/LLM 与专用检测器，体现了未来评测的方向：**通用语义推理 + 专项视觉测量 + 人类校准**。

29.6.3 EvalCrafter

EvalCrafter 以多维自动指标、prompt 分类和人评验证来比较开放域 T2V。它说明一个重要原则：指标集的价值不仅在总排名，更在于构建“模型-能力矩阵”，用于发现特定失败类型。

29.6.4 T2V-CompBench

T2V-CompBench 聚焦组合性，如属性绑定、空间关系、动作绑定和复杂组合。组合评测应特别警惕 bag-of-concepts 取巧：主体都出现不代表关系正确。

29.6.5 ChronoMagic-Bench

ChronoMagic 面向时间推移类视频，例如花开、融化、建造、腐烂等。此类任务需要判断过程速度、状态覆盖和时间方向，不能只检查首末帧。

29.7 人类偏好：为什么配对通常优于绝对评分

绝对 1–5 分受评审尺度影响大；配对任务只问 A 是否优于 B，认知负担较低。Bradley-Terry 模型：

$$P(A \succ B) = \sigma(\theta_A - \theta_B),$$

最大化：

$$\mathcal{L}(\theta) = \sum_{(A,B)} \left[y \log \sigma(\theta_A - \theta_B) + (1-y) \log \sigma(\theta_B - \theta_A) \right].$$

若有平局，可使用 Davidson 模型或将平局作为第三类。若比较多个模型和多个 prompt，应加入 prompt 难度与评审随机效应，而不是把所有 pair 混成独立伯努利。

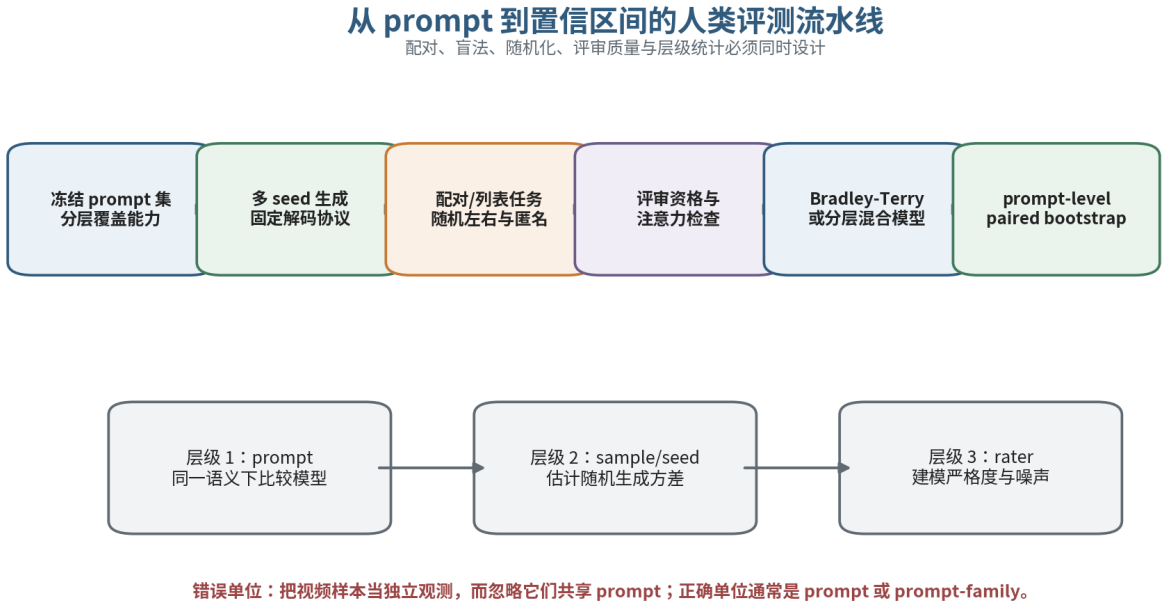

图 3: 人类评测与层级统计流水线

29.8 Paired bootstrap

对每个 prompt 得到差值：

$$d_q = \frac{1}{S} \sum_s \left(Y_{q,s,A} - Y_{q,s,B} \right).$$

从 prompt 集有放回采样 B 次：

$$\Delta^{(b)} = \frac{1}{Q} \sum_{q \in \mathcal{I}_b} d_q.$$

使用 $\{\Delta^{(b)}\}$ 的分位数给出置信区间。该方法保留同一 prompt 上的配对结构，通常比把所有视频独立 bootstrap 更合理。

29.8.1 多重比较

若比较 10 个模型共有 45 个 pair，直接使用 $p < 0.05$ 会产生大量偶然显著。可使用：

- Holm-Bonferroni 控制 family-wise error；
- Benjamini-Hochberg 控制 FDR；
- 预先指定少数主要比较；

- 报告 effect size 和 CI，而不是只报告星号。

29.9 推荐的“指标组合”

研究结论	最低证据组合
更清晰	VAE oracle + 感知质量 + 人评，排除锐化伪影
更符合文本	原子命题/组合 benchmark + 人类配对
运动更好	主体/相机分离的运动指标 + 事件正确性 + 人评
物理更真实	物理专项 benchmark + 轨迹/状态残差 + 专家审计
长视频更一致	分段质量曲线 + 身份/状态记忆 + 事件图 + 长程人评
世界模型更强	动作可控性 + 反事实分支 + rollout 稳定性 + 任务回报
后训练有效	reward、held-out judge、人评、多样性和成本 Pareto
更安全	分层红队 + subgroup + 漏报/误报 + provenance/响应演练

第 30 章 组合、时间、物理与因果：过程感知评测

“视频中包含 prompt 的关键词”只证明存在性，不证明过程正确。现代视频生成的核心难点是：同一实体在时间中保持身份，状态按照指定顺序变化，交互产生合理结果，且变化服从物理与常识。

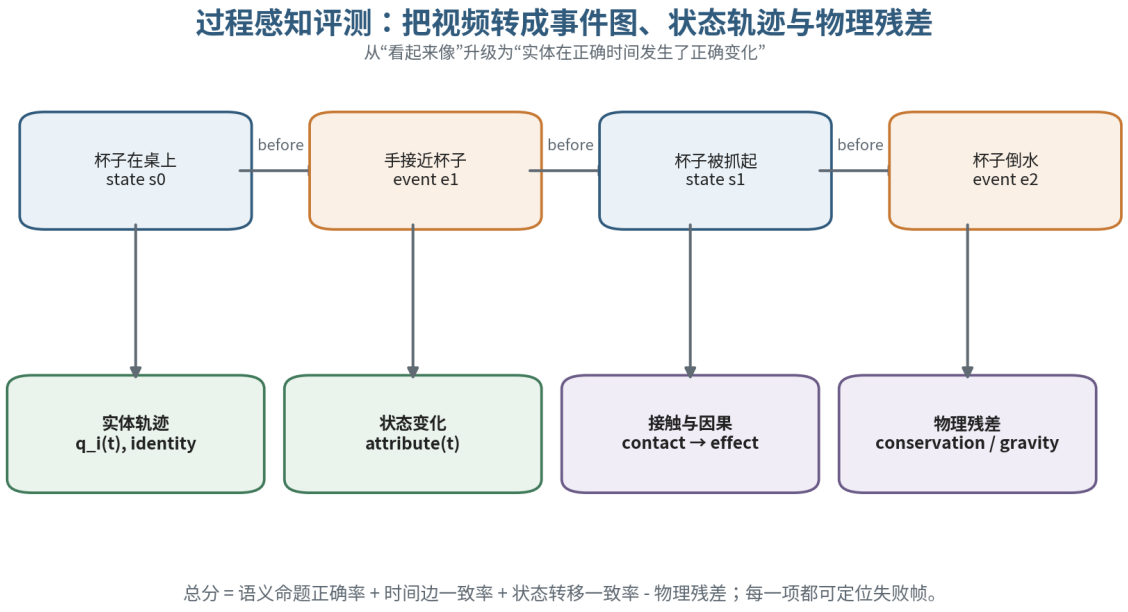

图 4: 事件图、状态轨迹与物理残差

30.1 从文本到事件图

定义事件图：

$$\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q).$$

节点可包括：

- 实体 o_i；
- 属性状态 a_i^t；
- 动作/事件 e_j；
- 场景或镜头 c_k。

边可包括：

- 时间顺序 $e_i \prec e_j$；
- 同时 $e_i \parallel e_j$；
- 因果 $e_i \to e_j$；
- 施事/受事 `agent(e,o)`、`patient(e,o)`；
- 空间 `left_of(o_i,o_j,t)`；
- 状态转移 $a_i^{t^-} \to a_i^{t^+}$。

生成视频通过 tracker、detector、VLM 和人工校验得到预测图 $\widehat{\mathcal{G}}_{\mathbf{x}}$。可计算节点和边的 precision/recall/F1：

$$F_1^{\text{edge}} = \frac{2P_{\text{edge}}R_{\text{edge}}}{P_{\text{edge}} + R_{\text{edge}}}.$$

事件图的优势是可解释：失败可以定位为“实体缺失”“施事交换”“顺序反转”“状态未保持”或“因果结果缺失”。

30.2 时间定位与顺序

对事件 e_j 预测起止区间 $[\hat{t}_j^s, \hat{t}_j^e]$。若有参考区间，可用 temporal IoU：

$$\mathrm{tIoU}(j) = \frac{|[t_j^s, t_j^e] \cap [\hat{t}_j^s, \hat{t}_j^e]|}{|[t_j^s, t_j^e] \cup [\hat{t}_j^s, \hat{t}_j^e]|}.$$

开放生成通常没有唯一参考时间，因此更常测相对顺序：

$$S_{\mathrm{order}} = \frac{1}{|\mathcal{E}_{\prec}|} \sum_{(i,j)\in\mathcal{E}_{\prec}} \mathbb{I}[\hat{t}_i^e < \hat{t}_j^s + \epsilon].$$

ϵ 允许动作重叠。应区分：

- 严格先后；
- 部分重叠；
- 必须同时；
- 条件触发；
- 周期性重复。

30.2.1 时间方向反事实

将视频倒放或交换事件段，理想顺序指标应明显下降。若不下降，说明指标只识别事件存在，不识别方向。

30.3 实体身份与对象恒存

设 tracker 得到对象 i 的 embedding $\mathbf{h}_i(t)$。身份一致性可用：

$$S_{\mathrm{id}} = \frac{1}{|\Omega_i|} \sum_{(t,t')\in\Omega_i} \cos(\mathbf{h}_i(t), \mathbf{h}_i(t')).$$

但人脸识别或通用 embedding 对遮挡、姿态和非人物物体可能失效。更完整的对象恒存需要：

1. 可见时身份一致；
2. 短暂遮挡后重新出现仍是同一对象；
3. 属性和状态按因果延续；
4. 不会无原因复制或消失；
5. 镜头切换后角色和道具保持绑定。

可以建立对象槽：

$$\mathbf{z}_i(t) = [\mathrm{appearance}, \mathrm{position}, \mathrm{state}, \mathrm{visibility}],$$

并使用 Hungarian matching 在帧间匹配。复制/消失惩罚：

$$L_{\mathrm{count}} = \sum_t |\hat{n}_i(t) - n_i^{\mathrm{expected}}(t)|.$$

30.4 属性绑定与组合泛化

例：红色立方体在蓝色球体左侧，随后立方体变绿。要分别测：

- 实体类别；
- 初始颜色绑定；
- 初始空间关系；
- 变化对象正确；
- 变化后颜色正确；
- 球体颜色未被误改；
- 立方体身份未替换。

定义每个原子条件 $c_j \in \{0, 1\}$，严格成功率：

$$S_{\text{strict}} = \prod_{j=1}^{J} c_j.$$

平均成功率 $J^{-1}\sum c_j$ 用于诊断，严格成功率用于衡量端到端完成。两者必须同时报告，否则模型可能在每个子项上略有能力，却几乎从未完整满足 prompt。

30.5 物理评测的三个层次

层次 A：视觉常识判断

VLM 判断“杯子落地后是否反弹”“水是否从容器中流出”。覆盖广，但受 judge 感知与常识偏差影响。

层次 B：可观测轨迹残差

对对象中心 $\mathbf{q}_t$：

$$\mathbf{v}_t = \frac{\mathbf{q}_{t+1} - \mathbf{q}_t}{\Delta t}, \qquad \mathbf{a}_t = \frac{\mathbf{q}_{t+1} - 2\mathbf{q}_t + \mathbf{q}_{t-1}}{\Delta t^2}.$$

自由落体的简化残差：

$$\delta_g = \frac{1}{T-2}\sum_t \|\mathbf{a}_t - \mathbf{g}_{\text{image}}(t)\|_2.$$

由于单目视频缺少尺度和相机姿态，$\mathbf{g}_{\text{image}}$ 只是投影近似。需要估计相机运动、深度或使用相对规律。

层次 C：显式状态与守恒

碰撞前后动量残差：

$$\delta_p = \left\| \sum_i m_i \mathbf{v}_i^{\text{before}} - \sum_i m_i \mathbf{v}_i^{\text{after}} \right\|.$$

机械能残差：

$$\delta_E = \left| \sum_i \left(\frac{1}{2} m_i \|\mathbf{v}_i\|^2 + m_i g h_i \right)_{t_1} - \sum_i \left(\frac{1}{2} m_i \|\mathbf{v}_i\|^2 + m_i g h_i \right)_{t_2} - W_{\text{diss}} \right|.$$

真实开放视频通常不知道质量、尺度、摩擦和外力，因此不应假装得到精确物理量。更实际的方法是：

- 使用相对排序，如重物与轻物的加速度关系；
- 检测明显穿透、无接触加速、无来源复制；
- 在合成/仿真 prompt 上获得可观测 ground truth；
- 用专家规则和 VLM 共同判断；
- 报告不可判定比例。

30.6 接触、因果与“视觉结果正确但机制错误”

若 prompt 要求“球撞倒积木”，成功至少需要：

1. 球运动；
2. 球与积木接触；
3. 接触后积木倒下；
4. 时间顺序合理；
5. 不接触时不能提前倒下。

定义接触信号 $C(t)$ 与效果变化 $E(t)$，因果时间差：

$$\Delta t_{CE} = \min\{t : E(t) = 1\} - \min\{t : C(t) = 1\}.$$

合理区间应满足 $0 \le \Delta t_{CE} \le \Delta_{\max}$。若积木在接触前倒下，语义结果可能看似正确，但因果错误。

30.6.1 反事实干预

构造相同场景但移除碰撞物：

$$q' = \mathrm{do}(\text{ball absent}).$$

理想模型在 q' 中不应生成积木被撞倒。世界模型研究应比较：

$$P(E \mid \mathrm{do}(C = 1)) - P(E \mid \mathrm{do}(C = 0)).$$

虽然纯 T2V 不是可识别因果模型，但成对干预能检测是否只复现表面共现。

30.7 专项 benchmark 解读

VideoPhy / VideoPhy-2

VideoPhy-2 以动作中心的 prompt 测语义遵循、物理常识和细粒度物理规则。其公开结果显示，即使领先模型在 hard 子集上的语义与物理联合成功率仍很低，特别容易违反质量和动量等守恒规律。它适合检验“动作是否在真实世界可成立”，但自动 evaluator 仍需要目标模型之外的人类校准。

PhyCoBench 与 T2VPhysBench

这类 benchmark 按物理原则构建 prompt，并使用 VLM、规则或专家标注评估。研究者应检查：

- prompt 是否唯一对应某条物理规律；
- 视觉观测是否足以判断；
- judge 是否把美观当物理；
- 复杂相机是否造成误判；
- 是否有同场景反事实对照。

T2VTextBench

动态文字生成需要同时满足字形正确、位置稳定、透视/遮挡合理和时间变化正确。普通 OCR 只在可读帧上工作，无法测文字漂移或前后内容变化。推荐报告：

$$S_{\text{text}} = S_{\text{OCR}} \cdot S_{\text{temporal}} \cdot S_{\text{placement}}.$$

WorldReasonBench、WorldBench 与 MIND 类评测

2026 年出现的世界推理基准开始将目标从“视觉质量”推进到多步世界状态、动作后果、反事实和长期一致性。它们对 T2V 很有启发，但“在世界推理 benchmark 得分高”仍不自动证明模型可用于机器人规划，因为还需要动作接口、闭环误差、状态可观测性和安全约束。

30.8 自动评测器本身也要做 benchmark

给定人工标签 y 和自动分数 r，至少报告：

- Pearson/Spearman/Kendall；
- 二分类 AUC、F1、balanced accuracy；
- ECE/Brier score；
- 各能力、时长、风格和模型家族的 subgroup 性能；
- 对帧打乱、倒放、复制、静止等人为扰动的响应；
- adversarial prompts；
- judge 版本更新后的 regression test。

Brier score：

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2.$$

评测器的训练集不能与被测模型输出、benchmark test 或后训练 reward 数据重叠，否则可能产生循环论证。

30.9 过程感知评分伪代码

```
# Pseudocode: process-aware evaluation
prompt_graph = parse_prompt_to_event_graph(prompt)
tracks = track_entities(video)
state_series = infer_attributes_and_events(video, tracks)
pred_graph = build_video_event_graph(tracks, state_series)

node_score = graph_node_f1(prompt_graph, pred_graph)
edge_score = temporal_causal_edge_f1(prompt_graph, pred_graph)
identity_score = object_persistence(tracks)
physics_residual = evaluate_physical_constraints(state_series)

score = (
    w_node * node_score
    + w_edge * edge_score
    + w_id * identity_score
    - w_phys * physics_residual
)
return score, localized_failures(pred_graph, prompt_graph)
```

关键输出不是总分，而是 `localized_failures`：失败实体、关系、时间段和证据帧。

30.10 本章自检

练习与自检

1. 为“一个人点燃蜡烛，蜡烛融化并逐渐变短”画事件图，列出至少 6 个原子条件。2. 为什么首末帧正确仍不足以证明过程正确？3. 单目生成视频中，动量守恒残差有哪些不可识别变量？4. 设计一对反事实 prompts 来检测“接触导致物体倒下”。5. 自动 judge 与人类高度相关但对帧倒放不敏感，这说明什么？

第 31 章 Benchmark 体系与 Meta-Evaluation

31.1 Benchmark 不只是 prompt 列表

完整 benchmark 定义为

$$\mathcal{B} = (\mathcal{P}, \mathcal{G}, \mathcal{E}, \Pi, \mathcal{H}),$$

其中：

- $\mathcal{P}$：prompt、图像、视频、动作等任务输入；
- $\mathcal{G}$：任务分组与难度层级；
- $\mathcal{E}$：自动 evaluator 集合；
- Π：生成与评价协议；
- $\mathcal{H}$：人工标注、锚定集和 meta-evaluation 证据。

只公开 prompt 而不公开 evaluator 版本、预处理、人工协议和生成配置，无法保证可重复。

31.2 VBench 与 VBench++

VBench 将视频生成质量分解为多个维度，典型包括：

- subject consistency；
- background consistency；
- temporal flickering；
- motion smoothness；
- dynamic degree；
- aesthetic quality；
- imaging quality；

- object class、multiple objects、human action；
- color、spatial relationship、scene、appearance style；
- temporal style、overall consistency。

其核心贡献不是某个单一 evaluator，而是建立“维度化评测”范式。VBench++ 进一步扩展到 T2V 与 I2V、更多 prompt 语言与可信性维度。

正确解读

VBench 总分可写为

$$S_{\text{VBench}} = \sum_k w_k s_k,$$

但不同维度的尺度、上限和可靠性并不相同。总分差异应配合雷达图和逐维置信区间；否则两个模型可能总分相同，一个偏静态美观，另一个偏动态但外观稍弱。

可能的 gaming

- 降低运动，提高 subject/background consistency；
- 生成中心构图和简单背景，提高 aesthetic；
- 避免多对象交互；
- 利用 evaluator 对特定风格偏好；
- prompt rewrite 增加 evaluator 易识别词。

因此 benchmark 应补充“难度保持”和“动态条件分层”。

31.3 EvalCrafter

EvalCrafter 以多维自动指标、人类偏好和综合分析评估大视频生成模型。其价值在于系统梳理：

- 视觉质量；
- 文本–视频一致性；
- 运动质量；
- 时间一致性；
- 人类主观判断。

教材式使用方式不是直接复制综合分数，而是把它当作“指标候选池”，再根据任务构念选择子集，并在自己的模型域上重新做 meta-evaluation。

31.4 T2V-CompBench：组合性不能由全局相似度代替

组合 benchmark 重点覆盖：

- 属性绑定；
- 多对象关系；
- 空间关系；
- 动作绑定；
- 动作交互；
- 数量；
- 复杂组合。

对于实体 o_i、属性 a_j，核心不是都出现，而是正确绑定：

$$(o_1, a_1), \quad (o_2, a_2)$$

不能被交换为

$$(o_1, a_2), \quad (o_2, a_1).$$

因此 evaluator 必须保持对象级对应，不能只做 bag-of-concepts。

31.5 DynamicEval：从“有运动”到“运动符合指令”

动态评价应区分：

- 主体运动；
- 相机运动；
- 场景内非刚体运动；
- 运动幅度；
- 运动方向；
- 运动复杂度；
- 运动与文本一致性。

DynamicEval 一类工作使用大规模 pairwise 人类标注和多模型视频，专门校准运动维度。其启示是：运动 evaluator 不应只从 optical flow magnitude 构造，而应学习或验证人类对“自然、充分、符合 prompt”的联合偏好。

31.6 物理 benchmark 家族

VideoPhy

聚焦常识物理一致性，通常要求评价器理解对象、动作和物理结果。适合回答“当前模型是否经常违反直觉物理”，不等同于精确动力学估计。

PhyGenBench

将 prompt 按物理规律和领域组织，强调系统化覆盖，例如力学、材料、流体、光学等。其优势是可按规律分析；限制是文本 prompt 与视觉可观测性之间可能存在歧义。

T2VPhysBench

进一步强调多种物理规律、人工协议和模型间比较。对每个规律应检查 evaluator 是否真正看视频，还是仅根据 prompt 判断最可能结果。

PhyCoBench

强调复杂物理组合和多阶段交互。组合物理的难点是错误可来自：

$$\text{对象识别} + \text{关系理解} + \text{事件顺序} + \text{物理响应}.$$

因此总错误率应做层级分解。

PhyWorldBench

通过基础物理、组合物理和反物理 prompt，测试模型是否真正具有稳定物理先验。反物理 prompt 特别重要：如果模型无条件纠正用户要求，说明指令遵循不足；如果模型轻易生成反物理但 evaluator 仍给高分，说明评价失效。需要同时标注“按指令生成”与“符合现实物理”两个维度。

31.7 WorldModelBench

世界模型 benchmark 通常不只关心清晰度，而关注：

- 对象和场景一致性；
- 物理与几何；
- 动作可控性；
- 时间演化；
- 世界状态违反；
- 可支持规划的真实性。

WorldModelBench 一类工作通过大规模人类标注训练专用 judge，并建立 world-model violation taxonomy。使用时必须注意：专用 judge 仍然继承训练模型和数据分布，不能替代闭环验证。

31.8 长视频 benchmark：局部好不等于全局好

SLVMEval

SLVMEval 是评测器的 meta-benchmark：从长视频构造受控高/低质量对，覆盖长达约三小时的内容，检验 evaluator 能否识别十类退化。它不直接宣称哪个生成模型最好，而是回答“你的长视频评测工具可靠吗”。

LongVQUBench

LongVQUBench 面向长程视频质量理解，覆盖超过 1200 个长视频和 1500 个问题，分为：

1. Local Quality Understanding（局部事件质量）；
2. Cross-event Quality Reasoning（跨事件推理）；
3. Global Quality Understanding（全局质量）。

并通过稀疏 needle distortion 测试模型能否在长上下文中找到局部异常。它揭示的关键问题是：输入支持长视频，不等于能进行长程质量归因。

LoCoT2V-Bench

长视频组合评测需要验证跨镜头身份、事件链、全局 prompt 和局部细节。适合使用层级 prompt：全局故事 + shot-level 条件 + cross-shot constraints。

31.9 Benchmark contamination 与 evaluator leakage

生成模型可能在训练中见过 benchmark prompt、参考视频或其近重复。evaluator 也可能在 benchmark 标注上训练。需区分：

- **generator contamination**：模型记忆 prompt/视频；
- **evaluator contamination**：judge 见过测试样本；
- **cross-contamination**：生成器与 evaluator 使用相同基础模型或训练数据，产生相关偏差。

检测策略：

1. prompt n-gram/embedding 检索；
2. 视频 perceptual hash 与 embedding 近邻；
3. 保留私有 holdout；
4. 用程序化新组合生成 prompt；
5. 测试同义改写、实体替换和关系反转；
6. 报告基础模型重叠。

31.10 难度自适应与饱和问题

当强模型在固定 benchmark 上接近饱和，增加更多相同难度 prompt 不能提高区分度。可使用 Item Response Theory（IRT）。对模型能力 θ_m、题目难度 b_i、区分度 a_i：

$$P(Y_{mi} = 1) = \sigma(a_i(\theta_m - b_i)).$$

优势：

- 估计 prompt 难度和区分度；
- 去除所有题目等权的假设；
- 构造不同能力区间的测试；
- 检测过易、过难或无区分题。

对多维能力，可使用

$$P(Y_{mi} = 1) = \sigma(\mathbf{a}_i^\top \theta_m - b_i).$$

31.11 Benchmark 的最小公开包

应包含：

```
benchmark/
  README.md
  LICENSE
  prompts.jsonl
  prompt_schema.json
  category_taxonomy.yaml
  generation_protocol.yaml
  evaluator_versions.yaml
  human_annotation_guideline.pdf
  human_anchor_set/
  meta_eval_pairs/
  scripts/
  expected_checksums.txt
```

其中 prompts.jsonl 每条至少包含：

```
{
  "prompt_id": "physics_collision_0042",
  "prompt": "A red ball rolls down ...",
  "category": ["physics", "collision", "event_order"],
  "atomic_claims": [
    {"id": "c1", "type": "entity", "text": "red ball exists"},
    {"id": "c2", "type": "before", "lhs": "roll", "rhs": "collision"},
    {"id": "c3", "type": "effect", "text": "blocks fall after impact"}
  ],
  "difficulty": null,
  "public": true
}
```

31.12 建议的 benchmark 选择

研究目标	主 benchmark	必须补充
通用短 T2V	VBench++ / EvalCrafter	人评、效率、prompt-level CI
组合指令	T2V-CompBench	原子命题失败分析
动态增强	DynamicEval 类协议	相机/主体运动分离
物理生成	VideoPhy + PhyGen/PhyWorld 类	estimator 校准、反物理对照
长视频	LoCoT2V + 自建层级集	SLVMEval-style meta-eval
世界模型	WorldModelBench	动作条件和闭环任务
视频奖励模型	VideoFeedback/VideoScore 类	OOD 模型与反 gaming 测试

核心结论

Benchmark 的价值不在于产生一个排行榜，而在于把研究假设转成可重复的测试。最好的 benchmark 不只区分模型，还能解释模型为何失败、评价器何时失效，以及改进是否具有外推性。

第 32 章　人类评价与统计推断

32.1 人类评价不是“找几个人看看”

视频主观评价至少涉及：

- 评价构念与 rubric；
- 单视频评分还是成对比较；
- 评审筛选、培训和质量控制；
- 视频播放界面、分辨率、循环方式和音频；
- 模型盲化与候选随机化；
- 样本量与统计功效；
- 评审相关性与 prompt 配对；
- 多重比较；
- 争议样本与开放意见。

32.2 Likert/MOS 与成对偏好

绝对评分

评审给 1–5 或 0–100 分：

$$y_{ij} = \mu_i + b_j + \epsilon_{ij},$$

其中 μ_i 是视频质量，b_j 是评审尺度偏差。优点是每次只看一个视频；缺点是不同评审的尺度校准困难。

Mean Opinion Score：

$$\mathrm{MOS}_i = \frac{1}{n_i} \sum_j y_{ij}.$$

Likert 是有序类别，不严格等距。小样本时仅报告均值和 t-test 可能不合适，可使用有序 logit/probit mixed model。

成对比较

给两个视频 A, B，选择更好或平局。优点是判断简单、相对尺度稳定；缺点是比较数随模型数增长：

$$\binom{M}{2} P$$

其中 M 为模型数、P 为 prompt 数。

32.3 Bradley–Terry 模型

给每个模型能力参数 θ_i，则 i 胜过 j 的概率为

$$P(i \succ j) = \frac{\exp(\theta_i)}{\exp(\theta_i) + \exp(\theta_j)} = \sigma(\theta_i - \theta_j).$$

若 w_{ij} 是 i 胜 j 次数，log-likelihood 为

$$\ell(\theta) = \sum_{i<j} \left[w_{ij} \log \sigma(\theta_i - \theta_j) + w_{ji} \log \sigma(\theta_j - \theta_i) \right].$$

由于加同一常数不改变概率，需设

$$\sum_i \theta_i = 0$$

或固定一个模型为零。

Prompt 与评审随机效应

更完整的 mixed BT：

$$\mathrm{logit}\, P(i \succ j \mid p, r) = (\theta_i - \theta_j) + (u_{ip} - u_{jp}) + b_r + c_{\text{position}},$$

其中：

- u_{ip}：模型在 prompt p 的相对表现；
- b_r：评审偏差；
- c_{position}：左右位置偏差。

这样可避免把所有比较当成独立同分布。

32.4 Thurstone–Mosteller 与 Elo

Thurstone 假设潜在质量带高斯噪声：

$$U_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

因此

$$P(i \succ j) = \Phi\left(\frac{\theta_i - \theta_j}{\sqrt{2}\sigma}\right).$$

Elo 是在线更新形式：

$$R_i' = R_i + K(S_i - E_i),$$

$$E_i = \frac{1}{1 + 10^{(R_j - R_i)/400}}.$$

Elo 适合连续比赛式排行榜，但更新顺序、K 和非平稳评审会影响结果；论文静态分析更适合直接拟合 BT/Thurstone 并报告不确定性。

32.5 平局与“不确定”

强迫评审在近似视频中二选一会引入噪声。Davidson tie model 可令

$$P(i \sim j) \propto \nu\sqrt{\lambda_i \lambda_j}, \qquad \lambda_i = e^{\theta_i}.$$

实践中可提供：

- A 更好；
- B 更好；
- 相近/无法判断；
- 两者都失败。

“都失败”与“相近”语义不同，不应合并。

32.6 多维人评

建议至少分开：

1. prompt adherence；
2. visual quality；
3. motion/temporal quality；
4. physical plausibility；
5. overall preference。

如果先问 overall，再问细分，可能产生 anchoring。可随机题序或让不同评审负责不同维度。对每个维度定义可观察 rubric，例如：

- 0：关键约束缺失或相反；
- 1：部分满足，存在明显错误；
- 2：主要满足，有轻微偏差；
- 3：完整、明确满足。

32.7 Inter-rater agreement

Cohen’s κ

两评审分类一致性：

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

p_o 为实际一致率，p_e 为随机期望一致率。

Fleiss' κ

扩展到多个评审。对高度不平衡标签，κ 可能出现 paradox，应同时报告原始一致率和类别分布。

Krippendorff's α

适用于缺失值和多种尺度：

$$\alpha = 1 - \frac{D_o}{D_e}.$$

D_o 为观测 disagreement，D_e 为随机 disagreement。需根据 nominal/ordinal/interval 选择距离函数。

ICC

连续评分可使用 Intraclass Correlation Coefficient，但要明确是 absolute agreement 还是 consistency，以及评审是随机效应还是固定效应。

32.8 配对 bootstrap

比较模型 A、B。对 prompt i，先聚合 seed 得

$$d_i = s_i^A - s_i^B.$$

总体差异：

$$\widehat{\Delta} = \frac{1}{P}\sum_{i=1}^{P} d_i.$$

配对 bootstrap 每次有放回采样 prompt 索引：

$$\widehat{\Delta}^{*(b)} = \frac{1}{P}\sum_{i \in I_b} d_i.$$

用分位数得到 95% CI。关键是**重采样 prompt，而不是把每个 seed 或每个评审当作完全独立样本**。

若 benchmark 有类别，可用 stratified bootstrap 保持类别比例。

32.9 层级 bootstrap

当数据层级为 prompt–seed–rater，可做：

1. 重采样 prompt；
2. 每个 prompt 内重采样 seed；
3. 每个视频内重采样 rater。

这比平铺所有评分更保守，也更符合数据生成过程。

伪代码：

```
for b in range(B):
    prompts_b = resample(prompts)
    diffs = []
    for p in prompts_b:
        seeds_b = resample(seeds[p])
        prompt_values = []
        for s in seeds_b:
```

```
            raters_b = resample(raters[p, s])
            prompt_values.append(mean(score_A - score_B over raters_b))
        diffs.append(mean(prompt_values))
    delta_b[b] = mean(diffs)
```

32.10 随机化检验

零假设下 A/B 标签可交换。对每个 prompt 随机翻转差值符号：

$$d_i^* = \xi_i d_i, \qquad \xi_i \in \{-1, +1\}.$$

计算置换分布，得到 p-value。该方法不要求差值正态，但仍依赖交换性。

32.11 效应量，而不仅是 p-value

均值差：

$$\Delta = \bar{s}_A - \bar{s}_B.$$

标准化配对效应量：

$$d_z = \frac{\bar{d}}{s_d}.$$

偏好概率本身也直观：

$$P(A \succ B) = \sigma(\theta_A - \theta_B).$$

“统计显著但只有 50.8% 偏好”可能没有实际价值；“55% 偏好但 CI 宽”说明样本不足。

32.12 样本量与统计功效

近似检测均值差 δ：

$$n \approx \left(\frac{(z_{1-\alpha/2} + z_{1-\beta})\sigma_d}{\delta} \right)^2.$$

其中：

- α：I 类错误；
- $1-\beta$：power；
- σ_d：prompt 配对差值标准差。

最可靠做法是先用小规模 pilot 估计 σ_d，再做 power analysis。prompt 多样性不足时，增加评审数不能弥补覆盖不足。

32.13 多重比较

比较 M 个模型有 $M(M-1)/2$ 对。若每对都用 0.05，家族错误率上升。可采用：

- Bonferroni：$\alpha' = \alpha/K$；
- Holm step-down；
- Benjamini–Hochberg 控制 FDR；
- 直接拟合全局 BT 并对参数差做 simultaneous intervals。

但多重比较修正不能解决 benchmark 选择和 evaluator 过拟合。

32.14 评审质量控制

- 明确入选条件和知情同意；
- 插入 gold/control 样本；
- 检查过快完成、恒定选择和位置偏差；
- 对失败 control 的评审预先定义剔除规则；
- 不能在看结果后选择剔除阈值；
- 报告每样本评审数和剔除比例；
- 对可能令人不适内容给予提示与退出机制。

32.15 人评报告模板

```
Task: pairwise preference, prompt adherence
Models: A/B, anonymized, left-right randomized
Prompts: 300, stratified across 6 categories
Seeds: 2 per model per prompt
Raters: 5 independent judgments per pair
Playback: native aspect ratio, 720p viewport, loop allowed, muted
Tie options: A / B / similar / both fail
Quality control: 10% gold pairs, pre-registered exclusion
Statistics: mixed-effects Bradley–Terry, prompt random effect
Uncertainty: prompt-cluster bootstrap, 10,000 replicates
Multiple comparison: Holm correction
Primary endpoint: P(A > B)
Secondary endpoints: category-wise preference, disagreement, failure tags
```

常见误区

将同一个视频的 20 个评审票当作 20 个独立视频，会严重夸大样本量。统计单位取决于外推目标：若要推广到新的 prompt，prompt 必须是主要 cluster；若只关心固定 prompt 集，则结论范围应明确限制在该集合。

练习与自检

假设 A 对 B 的 600 个判断来自 100 个 prompt，每个 prompt 3 个 seed，每个 pair 2 个评审。解释为什么二项检验把 600 当独立样本不正确；分别给出 prompt-level bootstrap、mixed BT 和按 prompt 多数票三种分析，并比较它们的假设。

第 33 章 长视频生成：定义、误差累积与评价对象

33.1 “长”不是固定秒数

长视频的难度由多个尺度共同决定：

$$\mathcal{L} = (T,\ f_{\text{fps}},\ N_{\text{shots}},\ H_{\text{event}}, H_{\text{memory}},\ C_{\text{control}}).$$

- T：帧数或总时长；
- f_{fps}：帧率；
- N_{shots}：镜头数；
- H_{event}：事件依赖跨度；
- H_{memory}：需要记住状态的最长距离；
- C_{control}：文本、动作和镜头控制复杂度。

一个 60 秒固定监控画面可能比 10 秒多角色叙事容易；一个 20 秒单镜头物理交互可能比 2 分钟无状态蒙太奇更难。因此应报告**时间长度、语义长度和状态长度**。

33.2 三类长视频任务

单镜头连续生成

要求身份、场景、相机和物理连续，不能通过切镜掩盖错误。适合测试动力学和记忆。

多镜头叙事生成

需要 storyboard、shot plan、跨镜头身份和事件因果。局部帧可以不连续，但全局语义必须一致。

动作条件交互 rollout

给定动作 a_t 或策略，模型持续预测环境。核心是可控性和闭环稳定，而非电影美学。

三类任务不应共用一个“long video score”。

33.3 自回归误差递推

设真实状态转移

$$s_{t+1} = F(s_t, a_t),$$

模型转移

$$\hat{s}_{t+1} = \widehat{F}(\hat{s}_t, a_t).$$

定义误差 $e_t = \|\hat{s}_t - s_t\|$。若真实/模型动力学在局部满足 Lipschitz 条件，可有

$$e_{t+1} \le L e_t + \delta_t,$$

其中 $\delta_t = \|\widehat{F}(s_t, a_t) - F(s_t, a_t)\|$ 是单步模型误差。递推：

$$e_T \le L^T e_0 + \sum_{k=0}^{T-1} L^{T-1-k} \delta_k.$$

- $L < 1$：误差趋于收缩；
- $L \approx 1$：误差近似线性积累；
- $L > 1$：误差可能指数放大。

真实视频动力学不是全局 Lipschitz 的，但这条式子直观解释了为什么长 rollout 会漂移。

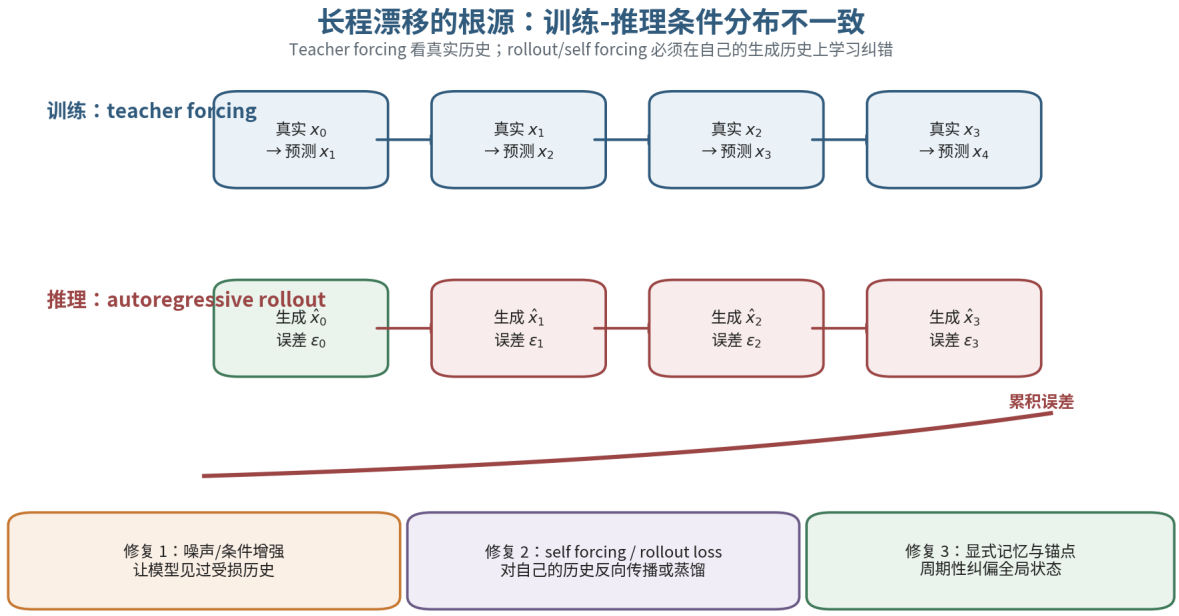

图 5: 长视频中的误差递推与多种漂移

33.4 Teacher forcing 与 exposure bias

训练时模型看到真实历史：

$$\mathcal{L}_{\mathrm{TF}} = -\sum_t \log p_\theta(x_t \mid x_{<t}^{\mathrm{real}}, y).$$

推理时历史来自模型：

$$x_t \sim p_\theta(x_t \mid \hat{x}_{<t}, y).$$

训练条件分布

$$p_{\mathrm{train}}(h_t)$$

与推理状态分布

$$p_\theta(h_t)$$

产生 covariate shift。Scheduled sampling、self-conditioning、on-policy rollout 和 causal post-training 都试图缩小这一差异。

33.5 扩散模型也存在暴露偏差

虽然一个短 clip 内是并行去噪，但长视频常通过窗口扩展：

$$x^{(k)} \sim p_\theta\left(x_{t_k:t_k+W} \mid x^{(<k)}, y\right).$$

上一窗口的生成结果成为下一窗口条件，所以仍有分布漂移。若只在真实前缀上训练，推理时的生成前缀同样是 OOD。

33.6 五类长程漂移

外观/身份漂移

人物脸、服装、物体颜色或结构缓慢变化。可测

$$D_{\mathrm{id}}(k) = 1 - \cos(e_0, e_k).$$

几何漂移

背景布局、房间拓扑、相机内参或物体相对位置不一致。

动力学漂移

速度、接触、重力或动作响应逐渐失真。

语义/叙事漂移

模型忘记早期设定，重复事件，角色目标改变或故事停滞。

质量漂移

画面逐渐模糊、饱和、纹理坍缩或运动冻结。

不同方法可能改善其中一种却恶化另一种。例如固定参考帧可以保持身份，却限制姿态变化或导致复制伪影。

33.7 长程状态与可恢复记忆

把历史压缩为记忆状态：

$$m_t = U_\eta(m_{t-1}, x_t, y),$$

$$x_{t+1:t+W} \sim p_\theta(\cdot \mid m_t, x_{t-R+1:t}, y).$$

理想记忆需满足：

- **充分性**：保留未来预测需要的信息；
- **紧凑性**：成本不随时长线性增长；
- **可更新性**：新事件能写入；
- **可检索性**：相关旧状态可被访问；
- **抗污染**：生成错误不会永久写入；
- **可解释性**：可诊断忘记了什么。

从信息论看，希望

$$I(m_t; x_{>t} \mid a_{\geq t}, y)$$

足够大，同时控制

$$I(m_t; x_{\leq t})$$

避免存储无关细节。

33.8 局部窗口平均的失效

长视频分成 K 个窗口，局部指标为 q_k。简单平均

$$\bar{q} = \frac{1}{K} \sum_k q_k$$

不检测窗口之间的身份或事件关系。应增加 cross-window 约束：

$$q_{\mathrm{cross}} = \frac{2}{K(K-1)} \sum_{i<j} \mathrm{Consist}(w_i, w_j).$$

全局质量可写为

$$Q_{\text{long}} = g\left(\{q_k\}, q_{\text{cross}}, q_{\text{story}}, q_{\text{state}}\right).$$

33.9 层级评价协议

局部层

每 2–8 秒窗口：清晰度、运动、闪烁、局部文本和物理。

中程层

跨相邻窗口：身份、场景、速度、相机、状态转移和边界伪影。

全局层

完整视频：故事目标、事件顺序、对象持久性、重复、节奏和结局。

稀疏异常层

检查是否存在少量灾难性失败：

$$Q_{\text{needle}} = \Pr(\text{detect sparse failure}).$$

该层与 LongVQUBench、SLVMEval 的长程质量理解思想一致。

33.10 长视频计算复杂度

若一次处理全部 token，full attention：

$$C_{\text{attn}} = O(N^2 d), \quad N \propto THW.$$

时长扩大 c 倍，attention 约扩大 c^2 倍。窗口方法令每窗口 token 为 N_w，窗口数 K：

$$C_{\text{local}} = O(KN_w^2 d).$$

若窗口 overlap 比例为 ρ，有效生成效率：

$$\eta_{\text{overlap}} = \frac{W - O}{W} = 1 - \rho.$$

记忆 cross-attention 若 M 个 memory tokens：

$$C_{\text{mem}} = O(KN_w M d).$$

因此长视频方法是在三者间权衡：

$$\text{全局一致性} \leftrightarrow \text{局部容量} \leftrightarrow \text{计算/存储成本}.$$

33.11 长视频中的帧率陷阱

“生成 60 秒” 需要同时说明：

- 原生生成 fps；
- 是否通过插帧提高 fps；
- 去重后的有效帧率；
- 是否有冻结段；
- 模型实际预测的帧数；
- 语义事件密度。

如果模型生成 2 fps 再插值到 24 fps，不能与原生 24 fps 的动态建模直接等价。

33.12 长视频的最低报告协议

```yaml
long_video:
  duration_sec: 120
  output_fps: 24
  native_generated_fps: 8
  interpolation: rife_v4
  native_frames: 960
  output_frames: 2880
  shot_count: 12
  generation_mode: hierarchical_storyboard
  window_frames: 64
  overlap_frames: 16
  memory_tokens: 256
  reset_policy: per_shot
  reference_refresh: true

metrics:
  local_window_sec: 4
  cross_window_lags_sec: [4, 16, 60]
  identity_drift: true
  scene_topology: true
  event_graph: true
  repetition: true
  failure_duration: true
  sparse_needle_audit: true
```

核心结论

长视频研究的基本单位不是“多生成了多少帧”，而是“在多长的依赖跨度上保持了哪些状态约束”。时长、镜头数、事件链、记忆跨度和原生帧率必须同时报告。

第 34 章　长视频方法谱系：窗口、记忆、自回归与因果生成

34.1 方法设计空间

长视频架构可抽象为

$$x_{1:T} = \mathrm{Compose}\left(G_\theta(\epsilon_k, c_k, m_k, x_{\mathrm{anchor}(k)})\right)_{k=1}^{K}.$$

关键选择：

- 窗口如何移动；
- 窗口是否同时去噪；
- 过去信息以像素、latent、KV、语义或状态保存；
- 训练是否看生成历史；
- 是否有全局计划；
- 误差何时重置或纠正。

34.2 Training-free noise rescheduling：FreeNoise

短视频扩散模型通常只在固定长度 W 上训练。FreeNoise 的代表思想是：在不训练或少训练条件下，对长序列噪声和 temporal attention 做重排，使局部窗口符合训练分布，同时在全局共享噪声相关性。

可抽象为：

$$\epsilon_{1:T} = \mathcal{R}(\epsilon^{(1)}_{1:W}, \dots, \epsilon^{(K)}_{1:W}),$$

其中 $\mathcal{R}$ 通过 shuffle/reuse 保持每个窗口的边际噪声模式。优势：

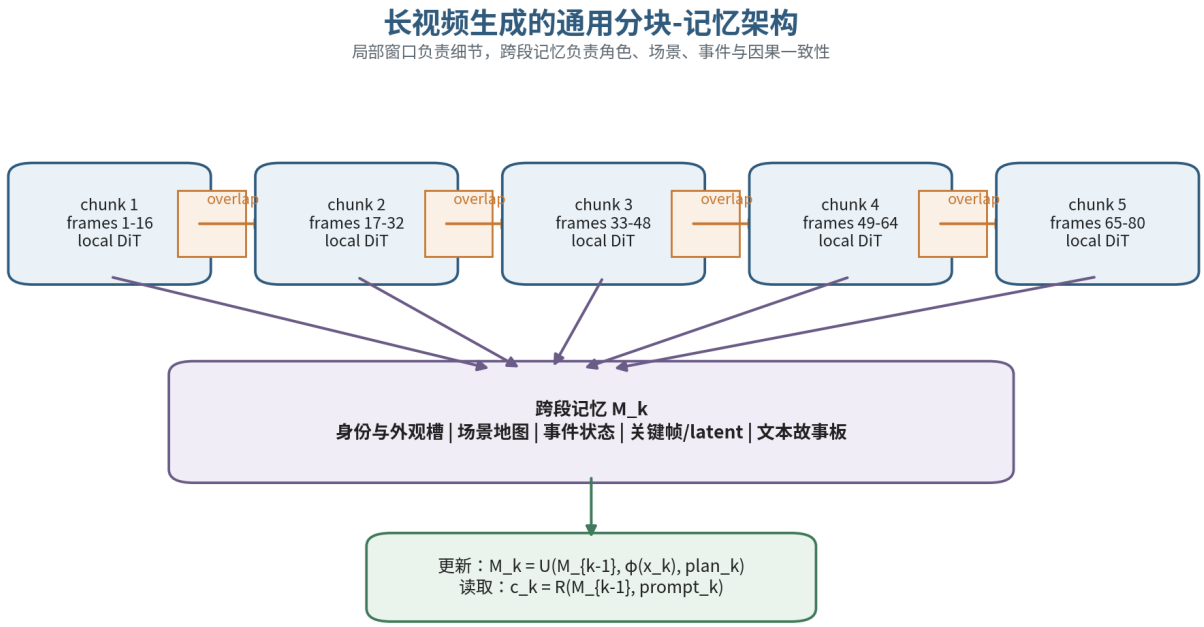

图 6: 长视频方法的四种架构路线

- 无需重新训练大模型；
- 可扩展预训练短视频模型；
- 易于原型验证。

限制：

- 不能创造模型未学到的长期动力学；
- 全局语义和身份仍可能漂移；
- 噪声一致不等于状态一致；
- 长度越远，条件约束越弱。

34.3 FIFO-Diffusion：队列式对角去噪

FIFO 类方法维护长度为 W 的 latent 队列。不同位置处于不同噪声水平，新噪声从一端进入，干净帧从另一端输出：

$$Q_k = [z_{k,\tau_1}, z_{k+1,\tau_2}, \dots, z_{k+W-1,\tau_W}],$$

每轮：

1. 对队列执行一组去噪更新；
2. 输出最干净位置；
3. 左移队列；
4. 注入新的高噪 latent。

它把二维“帧位置–噪声时间”组织成对角线。优点是连续流式输出、窗口成本固定；风险是队列边缘的上下文非对称、长期语义弱化和错误持续传播。

34.4 StreamingT2V：短期运动与长期外观分工

StreamingT2V 一类方法将：

- 短窗口模型负责局部运动；
- 条件增强/外观保持模块负责长期一致；
- 过去帧或锚点提供跨块条件。

统一写作

$$x_{k:k+W} \sim p_\theta\left(\cdot \mid x_{k-R:k}, A(x_{1:k}), y\right),$$

A 为长期外观聚合器。设计难点是避免 anchor 过强导致运动冻结，或过弱导致身份漂移。

34.5 ARLON：粗粒度自回归计划 + 细粒度扩散渲染

ARLON 代表一种分层思想：先自回归生成较低维、较粗的长程 latent/token 计划，再由视频扩散模型细化。概率分解：

$$p(x \mid y) = \sum_c p_\omega(c \mid y) p_\theta(x \mid c, y).$$

其中 c 可以具有长时序结构但较低空间带宽。它缓解：

- 直接在高维视频 latent 上做长自回归的成本；
- 纯扩散窗口缺乏长程计划的问题。

但会引入 planner–renderer interface error，与 Bernini 的语义计划诊断类似：

$$\mathcal{E} = \mathcal{E}_{\text{plan}} + \mathcal{E}_{\text{render}} + \mathcal{E}_{\text{interface}}.$$

34.6 Ca2-VDM：自回归因果视频扩散

Causal autoregressive video diffusion 将过去 clip 作为条件，生成下一 clip：

$$p(x_{1:K} \mid y) = \prod_{k=1}^{K} p_\theta(x^{(k)} \mid x^{(<k)}, y).$$

训练可采用 causal attention、context dropout 和不同历史长度。优势是自然支持 streaming；限制是 exposure bias、生成历史污染和长期状态压缩。

34.7 Self-Forcing：在模型自己的历史上训练

Self-forcing 的核心是让训练条件更接近推理：

1. 使用当前模型 rollout 生成历史；
2. 在生成历史条件下学习下一块；
3. 通过 stop-gradient、teacher/student 或分阶段稳定训练。

目标可写为

$$\mathcal{L}_{\text{SF}} = \mathbb{E}_{\widehat{h} \sim p_\theta} \ell\left(G_\theta(\widehat{h}), x_{\text{target}}\right).$$

与 teacher forcing 的区别是历史分布来自 p_θ。Self-Forcing++ 等工作进一步探索很长 rollout 和效率优化，并报告远超教师训练 horizon 的生成长度。阅读这类结果时应核对：

- “连续生成”是否单镜头；
- 原生 fps 和插帧；
- 是否使用参考刷新；
- 质量在哪个时长开始下降；
- 是否有人工挑选；
- 训练和评测是否使用相同域。

34.8 VideoSSM：状态空间模型作为长期记忆

状态空间模型递推：

$$h_{t+1} = A_t h_t + B_t u_t,$$

$$y_t = C_t h_t + D_t u_t.$$

在视频中，u_t 可以是帧/patch feature，h_t 是长期状态。线性或选择性扫描可将序列复杂度降至近线性。VideoSSM 类方法尝试将 SSM 的长程效率与视频生成结合。

优势：

- 状态大小固定；
- 长序列计算更友好；
- 流式推理自然。

难点：

- 高维视觉细节能否被固定状态保留；
- 状态污染和不可逆遗忘；
- 与 2D/3D 空间建模的耦合；
- 并行训练与严格因果推理的一致性。

34.9 Causal Forcing 与架构间隙

长视频生成不只存在 training distribution gap，还存在 **architecture gap**：训练时双向/全局去噪结构可能依赖未来 token，推理改为因果滚动时信息结构发生变化。

设训练 operator 为

$$G_{\text{bi}}(z_{1:T}),$$

推理 operator 为

$$G_{\text{causal}}(z_t, h_{t-1}).$$

即使参数相同，两者可表示函数族也不同。Causal Forcing 类研究从可逆性、信息注入和因果结构分析这一差距，并设计训练使因果推理时的状态映射更稳定。

34.10 Storyboard–shot–frame 三层生成

多镜头长视频更适合层级分解：

$$y \xrightarrow{\text{story planner}} S_{1:M} \xrightarrow{\text{shot planner}} C_{1:K} \xrightarrow{\text{renderer}} x_{1:T}.$$

- S_m：故事事件、角色目标和因果；
- C_k：镜头 prompt、时长、相机、角色状态和连接方式；
- x：像素视频。

关键状态表：

```
{
  "character_A": {
    "identity_ref": "ref_A.png",
    "clothes": "blue coat",
    "location": "station platform",
    "goal": "board the train",
    "carried_objects": ["red suitcase"],
    "last_seen_shot": 4
  }
}
```

每个镜头生成后，视觉理解模型更新状态，再交给下一镜头。该流程提高可控性，但理解错误也可能写入记忆，因此需要置信度和回滚。

34.11 Memory 的五种载体

载体	优势	风险
像素/关键帧	直接、身份信息丰富	成本高、复制和冻结
VAE latent	与生成器接口自然	依赖 VAE、难解释
Transformer KV	计算复用	随长度增长、旧错误难清除
语义 token/状态表	紧凑、可编辑	丢失细节、planner 错误
3D/4D 场景表示	几何稳定、可重投影	重建困难、动态物体复杂

下一代系统可能使用混合记忆：

$$m_t = (m_t^{\text{semantic}}, m_t^{\text{identity}}, m_t^{\text{geometry}}, m_t^{\text{recent}}).$$

34.12 长视频训练课程

建议从短到长，同时控制状态难度：

1. 短 clip 重建和无条件动力学；
2. 真实历史条件下一块预测；
3. 轻度污染历史；
4. 自生成历史；
5. 长 rollout；
6. 跨镜头故事与状态更新；
7. OOD 动作和反事实。

训练采样长度 L 可采用 curriculum：

$$P_t(L) \propto \exp(-\lambda_t L),$$

随训练推进减小 λ_t，增加长样本比例。

34.13 长程稳定性的消融矩阵

至少比较：

变量	设置
历史来源	real / generated / mixed
memory	none / keyframe / semantic / latent / hybrid
horizon	1x / 2x / 4x / 8x train horizon
reset	never / per shot / confidence-triggered
anchor strength	low / medium / high
overlap	0 / 25% / 50%
evaluator	local / cross-window / global / closed-loop

并画随时间的曲线：

$$q(t), \quad D_{\text{id}}(t), \quad E_{\text{geom}}(t), \quad R_{\text{freeze}}(t).$$

只在最终时长报告一个分数无法揭示失稳起点。

34.14 一个通用流式生成伪代码

```
memory = init_memory(prompt, references)
recent = []
outputs = []

for k in range(num_chunks):
    condition = build_condition(
        prompt=prompt,
```

```
        storyboard=storyboard[k],
        memory=memory,
        recent_frames=recent,
    )
    chunk = sample_video_chunk(condition, seed=base_seed + k)
    diagnostics = evaluate_chunk_and_boundary(chunk, recent, memory)

    if diagnostics["catastrophic_failure"]:
        chunk = repair_or_resample(chunk, condition, diagnostics)

    outputs.append(remove_overlap(chunk))
    recent = update_recent(chunk)
    memory = update_memory(
        old_memory=memory,
        observed_chunk=chunk,
        confidence=diagnostics["state_confidence"],
    )

return concatenate(outputs)
```

这个流程暴露了研究接口：`build_condition`、`update_memory`、`repair_or_resample` 和 `state_confidence` 都可成为独立论文问题。

常见误区

Training-free 长视频扩展能改变噪声组织和局部上下文，却不能保证模型获得新的长期因果能力。生成更长与建模更长不是同义词；前者是输出长度，后者是可验证的依赖跨度。

练习与自检

选择一个短视频底座，实现三种扩展：重叠窗口、关键帧锚定、语义状态记忆。固定总 NFE 和输出长度，比较身份漂移、运动冻结、边界伪影、全局 prompt 和峰值显存。要求画出每个指标随时间的曲线，而不仅是终点平均。

第 35 章　多镜头故事、MLLM 规划与长期一致性

长单镜头强调连续动力学，多镜头故事强调**叙事状态与视觉身份跨镜头保持**。电影式视频通常包含 shot、scene、sequence 三层：

- **shot**：一次连续拍摄/生成片段，无硬切；
- **scene**：同一地点或连续事件的一组 shot；
- **sequence**：实现更高层叙事目的的一组 scene。

35.1 故事生成的分层概率模型

设故事文本为 q，镜头计划为 $\mathbf{p}_{1:K}$，视频镜头为 $\mathbf{x}_{1:K}$：

$$p(\mathbf{x}_{1:K}, \mathbf{p}_{1:K} \mid q) = p_\omega(\mathbf{p}_{1:K} \mid q) \prod_{k=1}^{K} p_\theta(\mathbf{x}_k \mid \mathbf{p}_k, \mathbf{M}_{k-1}, q).$$

Planner 负责：

- 将长 prompt 拆成事件；
- 决定镜头边界、时长和相机；
- 维护角色/道具/地点状态；
- 选择参考图或关键帧；
- 检查叙事约束；
- 在生成失败时局部重规划。

Renderer 负责每个镜头的像素合成。Bernini 的 latent semantic planning 可视为更连续、更接近视觉空间的 planner-renderer 接口；故事系统还需要跨镜头状态机。

35.2 故事状态

定义：

$$\mathbf{S}_k = (\mathcal{C}_k, \mathcal{O}_k, \mathcal{L}_k, \mathcal{R}_k, \mathcal{E}_{\le k}, \mathcal{G}_k),$$

其中：

- $\mathcal{C}_k$：角色及外观；
- $\mathcal{O}_k$：道具和持有关系；
- $\mathcal{L}_k$：地点与空间布局；
- $\mathcal{R}_k$：角色关系/目标；
- $\mathcal{E}_{\le k}$：已完成事件；
- $\mathcal{G}_k$：未完成叙事目标。

状态转移：

$$\mathbf{S}_{k+1} = F_\omega(\mathbf{S}_k, \mathbf{p}_k, \hat{\mathbf{e}}_k),$$

$\hat{\mathbf{e}}_k$ 是从实际生成镜头提取的事件。不能只用计划事件更新，因为 renderer 可能没有成功生成；否则 planner 会误以为剧情已推进。

35.3 Open-loop 与 closed-loop story planning

Open-loop

一次性生成全部 storyboard：

$$\mathbf{p}_{1:K} = P_\omega(q).$$

优点：全局结构清晰、可并行生成。缺点：无法根据实际镜头失败修正。

Closed-loop

每生成一段，VLM/critic 读取结果：

$$\hat{\mathbf{e}}_k = V(\mathbf{x}_k),$$

$$\mathbf{p}_{k+1} = P_\omega(q, \mathbf{S}_k, \hat{\mathbf{e}}_k).$$

闭环更稳健，但会引入 judge 错误、额外延迟和误差级联。需要保存 planned state 与 observed state 的差异：

$$\Delta\mathbf{S}_k = d(\mathbf{S}_k^{\text{plan}}, \mathbf{S}_k^{\text{obs}}).$$

35.4 角色一致性不是单一人脸相似度

跨镜头角色一致性包含：

- 脸和身体身份；
- 发型、年龄、服装；
- 动作习惯与角色属性；
- 视角变化下的外观；
- 多角色不互换；
- 角色与名字/关系绑定。

角色 bank：

$$\mathbf{B}_i = [\mathbf{e}_i^{\text{face}}, \mathbf{e}_i^{\text{body}}, \mathbf{e}_i^{\text{style}}, \mathbf{t}_i^{\text{description}}].$$

每个镜头从 bank 检索条件。强制参考过强会抑制姿态变化；过弱会身份漂移。因此需要可学习门控或按层注入：低层保持纹理，高层允许动作和视角变化。

35.5 道具和地点的关系记忆

例如钥匙从 A 转移到 B：

$$\text{owner}(key, k) = B.$$

后续镜头若 A 再拿出同一钥匙而没有转移事件，就违反状态。故事评测应建立知识图谱：

$$(A, \text{gives}, key, B, t_3), \qquad (B, \text{owns}, key, t > t_3).$$

地点一致性也不能只比较背景 embedding。应维护：门、窗、桌子和出口的相对布局；镜头切换可改变视角，不应改变拓扑。

35.6 VideoGen-of-Thought 与“先思考再渲染”

这类方法将生成分成中间推理/规划步骤，例如：

1. 解析人物与事件；
2. 生成故事板/关键帧；
3. 推断中间状态；
4. 对每段生成运动；
5. 合并和修复。

其一般形式：

$$p(\mathbf{x} \mid q) = \int p_\theta(\mathbf{x} \mid \mathbf{h}, q) p_\omega(\mathbf{h} \mid q) d\mathbf{h}.$$

$\mathbf{h}$ 可以是文字 CoT、图像关键帧、布局、轨迹、ViT semantic plan。必须通过以下实验证明中间变量有效：

- no-plan；
- predicted-plan；
- oracle-plan；
- shuffled-plan；
- wrong-plan；
- plan dropout；
- 控制计划带宽与额外算力。

若 shuffled-plan 不影响结果，renderer 可能忽略计划。

35.7 StoryMem、OneStory 与专用数据

长故事需要包含角色复现、地点回访、道具状态和多事件的训练数据。普通短 clip 数据即使数量巨大，也很少提供同一角色跨分钟的干净 supervision。数据建设应保存：

- parent video 与 shot 层级；
- 角色 track 和跨 shot ID；
- shot caption 与全局 synopsis；
- 事件图；
- 道具状态；
- 地点关系；
- 镜头类型与切换；
- 音频/对白（若使用）；
- 版权和人物授权。

StoryMem、OneStory 等方向体现了显式记忆或专用长故事数据的需求。研究时应检查：提升来自模型机制还是数据中更强的角色重复 supervision。

35.8 故事评测

35.8.1 三层评分

$$S_{\text{story}} = w_1 S_{\text{shot}} + w_2 S_{\text{cross-shot}} + w_3 S_{\text{narrative}}.$$

- S_{shot}：单镜头质量和 prompt 遵循；
- $S_{\text{cross-shot}}$：角色、道具、地点和风格一致；
- $S_{\text{narrative}}$：事件顺序、因果和目标完成。

35.8.2 必须报告错误恢复

给中间镜头注入失败或替换为不一致镜头，系统是否能：

- 检测失败；
- 局部重生成；
- 更新故事状态；
- 避免后续继续沿错误状态生成?

这比在无干扰条件下生成一条成功故事更接近生产系统。

35.8.3 长文本不等于长故事

复杂 prompt 可能被模型压缩成一个场景。应测事件覆盖率：

$$S_{\text{coverage}} = \frac{|\hat{\mathcal{E}} \cap \mathcal{E}_q|}{|\mathcal{E}_q|},$$

以及重复率：

$$S_{\text{repeat}} = \frac{\#\{\text{重复且非计划事件}\}}{|\hat{\mathcal{E}}|}.$$

35.9 音频、对白和口型（扩展）

多镜头电影式生成还可能联合音频：

$$p(\mathbf{x}, \mathbf{a} \mid q) = p(\mathbf{x} \mid q) p(\mathbf{a} \mid \mathbf{x}, q)$$

或联合建模：

$$p(\mathbf{x}, \mathbf{a} \mid q).$$

应分别测：

- 语义一致；
- 音画事件同步；
- 说话人身份与声纹；
- 口型同步；
- 场景声连续；
- 镜头切换时音频是否合理延续。

不能用纯视频质量指标评价联合音视频系统。

35.10 研究建议：从结构化 planner 开始

资源有限时，不必一开始训练 7B MLLM + 14B Renderer。可按以下阶梯：

1. 冻结开源视频模型；
2. 用 LLM 将长故事转为 JSON storyboard；
3. 用现有人物参考/LoRA 保持角色；
4. 维护显式状态表；
5. 生成后用 VLM 更新 observed state；

6. 实现局部重规划；
7. 再将结构状态投影为连续 semantic token；
8. 最后研究端到端 planner-renderer 联合训练。

这一顺序能先验证“闭环状态是否提高长程故事”，再投入大规模训练。

第 36 章　从视频生成器到世界模型

36.1 “世界模型”至少有五个层级

“视频生成模型具有世界知识”不等于“它是可用于决策的世界模型”。可以建立能力阶梯：

1. **视觉生成器**：从文本/图像生成逼真视频；
2. **条件未来预测器**：给定过去，预测可能未来；
3. **动作条件模拟器**：未来受动作控制；
4. **可规划世界模型**：rollout 能正确比较动作后果；
5. **闭环交互模型**：在策略反馈下长期稳定，并能迁移到真实环境。

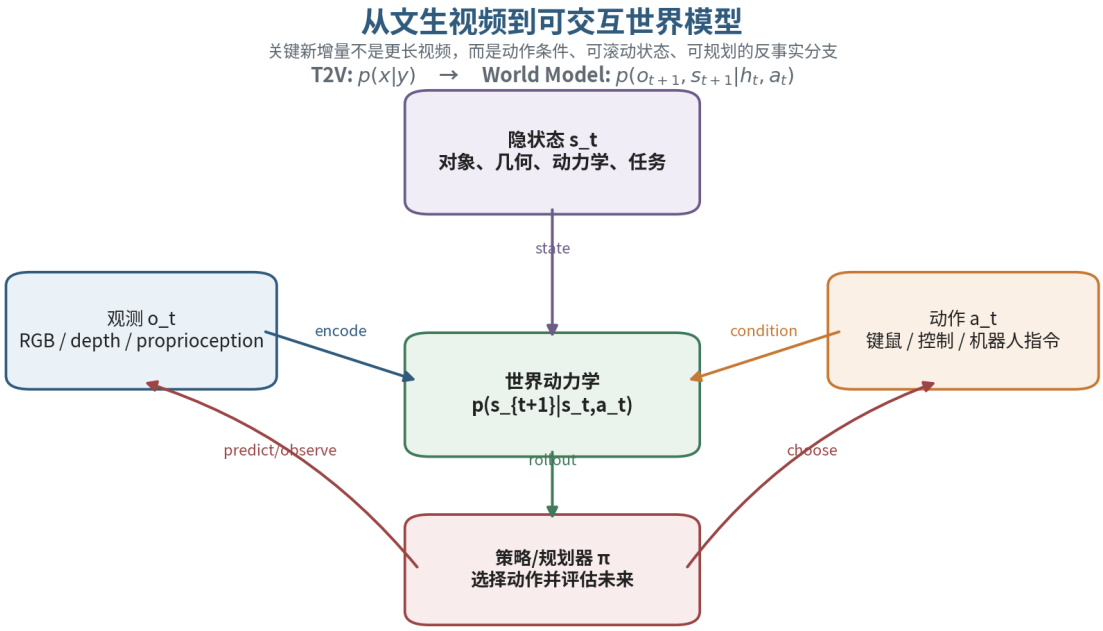

图 7: 从视觉生成到闭环世界模型的能力阶梯

越往上，视觉美学的重要性相对下降，动作响应、状态可辨识性、校准和闭环效用的重要性上升。

36.2 POMDP 统一表述

部分可观测马尔可夫决策过程：

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, O, R, \gamma).$$

- $s_t \in \mathcal{S}$：真实环境状态；
- $a_t \in \mathcal{A}$：动作；
- $o_t \in \mathcal{O}$：视觉/传感观测；
- $P(s_{t+1} \mid s_t, a_t)$：动力学；
- $O(o_t \mid s_t)$：观测模型；
- $R(s_t, a_t)$：奖励。

视频世界模型通常只能看到 $o_{\le t}$，需维护 belief：

$$b_t(s) = p(s_t = s \mid o_{\le t}, a_{<t}).$$

更新：

$$b_{t+1} \propto O(o_{t+1} \mid s_{t+1}) \int P(s_{t+1} \mid s_t, a_t) b_t(s_t) ds_t.$$

神经世界模型用 latent h_t 近似 belief：

$$h_t = U_\eta(h_{t-1}, o_t, a_{t-1}).$$

36.3 三个概率模型不能混淆

文本到视频

$$p(x_{1:T} \mid y).$$

文本描述结果，但通常不提供逐步动作。

视频预测

$$p(x_{t+1:T} \mid x_{1:t}).$$

预测自然演化，可能是多模态未来。

动作条件世界模型

$$p(x_{t+1:T} \mid x_{1:t}, a_{t:T-1}).$$

动作必须对结果具有可识别、可组合和正确方向的影响。一个优秀 T2V 模型可在第一项很强，却在第三项失败。

36.4 Latent dynamics 与 observation decoder

经典 world model 分解：

$$z_t = E(o_t),$$

$$p_\theta(z_{t+1} \mid z_t, a_t),$$

$$p_\psi(o_t \mid z_t).$$

训练目标可能包括：

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}} + \lambda_{\text{reward}}\mathcal{L}_r + \lambda_{\text{term}}\mathcal{L}_{\text{terminal}}.$$

视频扩散世界模型则可直接对未来 latent 分布建模：

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\|v_\theta(z_\tau, \tau, c_{\text{history}}, a) - u_\tau\|^2\right].$$

36.5 动作表示是核心瓶颈

动作可能是：

- 键盘/手柄离散动作；
- 机器人连续关节控制；
- 末端执行器位姿；
- 自然语言动作；
- latent action；
- 高层技能 token。

离散动作：

$$a_t \in \{1, \dots, K\}.$$

连续动作：

$$a_t \in \mathbb{R}^{d_a}.$$

latent action model 从相邻观测推断动作：

$$\hat{a}_t = A_\omega(o_t, o_{t+1}).$$

它允许从无动作标签视频学习“可控变化”，但 latent action 可能混合：

- agent action；
- 相机运动；
- 环境随机性；
- 不可控对象行为。

若 latent action 不可辨识，模型可能能重建，却不能被策略可靠控制。

36.6 Genie：从无标签视频学习可玩环境

Genie 的代表性结构包括：

1. spatiotemporal tokenizer，将视频离散化；
2. latent action model，从帧间变化学习动作 token；
3. autoregressive dynamics model，预测下一视觉 token。

可写为

$$u_{1:T} = Q(x_{1:T}),$$

$$a_t^{\text{latent}} = A(u_t, u_{t+1}),$$

$$p(u_{t+1} \mid u_{\le t}, a_t^{\text{latent}}).$$

其意义不是单纯生成游戏视频，而是从大量无动作标注视频中发现交互控制轴。局限是 latent action 的语义与真实控制器未必对齐。

36.7 GameNGen：动作条件实时游戏模拟

GameNGen 展示了扩散模型可以在动作条件下实时模拟游戏视觉。典型训练流程：

1. 用 RL agent 收集游戏 trajectory；
2. 训练生成模型预测下一帧/短未来；
3. 条件包含过去帧和动作；
4. 通过噪声增强等方式提高对自身历史的鲁棒性。

论文报告在特定游戏中达到实时帧率和较高视觉保真，但其结论范围是特定封闭环境，不能直接外推到开放世界或真实物理。

36.8 GameGen-X：从单环境走向多游戏基础模型

GameGen-X 扩展到大量游戏视频和多种交互场景，采用基础预训练和 instruction tuning。关键研究问题从“能否模拟一个游戏”转为：

- 动作空间如何统一；
- 不同游戏相机与动力学如何共享；
- 视觉风格是否与动力学解耦；
- 新游戏的 few-shot 适配；
- 长期交互稳定性。

多域训练可能提高视觉泛化，但也可能平均化精确动力学。

36.9 Cosmos 系列：Physical AI 的世界基础模型

Cosmos 系列将大规模视频生成、物理理解、推理和动作建模整合为 Physical AI 基础模型。Cosmos-Predict 类模型强调未来预测；Reason 类模型强调物理和具身推理；Transfer 类模型使用结构化控制进行世界生成。

Cosmos 3 进一步提出统一处理和生成 language、image、video、audio 与 action 的 omnimodal mixture-of-transformers 架构。其研究意义是：

$$\text{understanding} + \text{generation} + \text{action}$$

可能共享统一序列接口。但“统一模态”仍不自动保证精确闭环控制；需要独立验证动作因果、时延、校准和真实策略收益。

36.10 机器人世界模型与 synthetic rollout

机器人应用中，世界模型可以：

- 预测候选动作后果；
- 生成训练策略的合成数据；
- 做 counterfactual augmentation；
- 评估危险动作；
- 将语言目标转为视觉子目标。

模型预测控制（MPC）：

$$\mathbf{a}^*_{t:t+H-1} = \arg\max_{\mathbf{a}} \mathbb{E}_{\hat{p}_\theta}\left[\sum_{k=0}^{H-1} \gamma^k r(\hat{s}_{t+k}, a_{t+k})\right].$$

只执行第一步，再观测并重规划。世界模型误差会导致 model exploitation：规划器寻找模型误差中的高奖励轨迹。需使用：

- ensemble uncertainty；
- pessimistic value；
- uncertainty penalty；
- real-environment validation；
- short receding horizon；
- action constraints。

36.11 Open-loop 与 closed-loop 指标

Open-loop prediction

给定真实历史和动作，预测未来：

$$\hat{x}_{t+1:t+H} \sim G(x_{\le t}, a_{t:t+H-1}).$$

测视觉、状态、动作响应和分布校准。

Closed-loop rollout

模型输出进入下一步条件，或策略根据模拟输出选择动作。测：

$$J_{\text{model}}(\pi), \qquad J_{\text{real}}(\pi).$$

关键 world-model validity：

$$\Delta J = |J_{\text{model}}(\pi) - J_{\text{real}}(\pi)|.$$

还应测策略排序保持：

$$\mathrm{RankCorr}\left(\{J_{\mathrm{model}}(\pi_i)\}, \{J_{\mathrm{real}}(\pi_i)\}\right).$$

若绝对回报偏差大但策略排序正确，模型仍可能适合筛选；反之视觉逼真但排序错误，不适合规划。

36.12 Calibration 与多模态未来

世界模型应表达不确定性。给事件 E 的预测概率 p，真实频率应近似 p：

$$P(E \mid \hat{p}(E) = p) \approx p.$$

多样本预测：

$$\hat{x}^{(k)} \sim p_\theta(x_{t+1:T} \mid h_t, a), \quad k = 1, \dots, K.$$

覆盖率：

$$\mathrm{Coverage} = P\left(x^* \in \mathcal{C}(\hat{x}^{(1:K)})\right).$$

仅用 best-of-K 误差会随 K 自然降低，必须固定样本预算并同时报告多样性和概率质量。

36.13 反事实与可干预性

世界模型应能回答：保持历史不变，仅改变动作 a_t，未来如何改变？

$$\Delta_a = d\left(G(h_t, a), G(h_t, a')\right).$$

理想模型满足：

- 对无关动作维度具有不变性；
- 对关键动作维度有正确响应；
- 改变局部动作不会无理由重画整个世界；
- 结果差异符合因果结构。

可用 action swap 测试：

$$(h, a, x^+), \quad (h, a', x^-).$$

评价器检查变化是否仅发生在预期对象和时间。

36.14 世界模型评价卡

维度	例子
视觉	清晰、无闪烁、身份一致
动力学	状态转移、接触、物体持久性
动作	可控、方向正确、幅度单调
长程	rollout 稳定、记忆、无坍缩
不确定性	calibration、mode coverage
规划	policy ranking、MPC return
迁移	新场景、新对象、新动作
安全	OOD 检测、失败时保守
效率	latency、RTF、部署成本

核心结论

一个模型是否是“世界模型”，不能由视频是否逼真决定。最强证据是：动作干预产生正确后果，长 rollout 保持状态，模型不确定性可校准，并且模型中的策略选择能够预测真实环境中的策略收益。

第 37 章　物理、组合与因果生成的建模路线

37.1 失败根源：像素建模不等于状态建模

视频生成器学习

$$p(x_{1:T} \mid c),$$

但物理规律作用于隐状态

$$s_t = (q_t, \dot{q}_t, m, \text{shape}, \text{contact}, ...).$$

像素是渲染结果：

$$x_t = R(s_t, \ell_t, c_t),$$

其中 ℓ_t 是光照、c_t 是相机。若模型没有显式或隐式分离状态与渲染，可能通过纹理相关性产生逼真画面，却不保持动力学。

37.2 结构化条件接口

把自然语言解析为结构化控制：

$$crightarrow(\mathcal{O}, \mathcal{R}, \mathcal{A}, \mathcal{T}, \mathcal{K}),$$

- $\mathcal{O}$：对象与属性；
- $\mathcal{R}$：空间和交互关系；
- $\mathcal{A}$：动作；
- $\mathcal{T}$：事件时间图；
- $\mathcal{K}$：相机和场景约束。

生成器条件：

$$p_\theta(x \mid y, c_{\text{layout}}, c_{\text{traj}}, c_{\text{depth}}, c_{\text{event}}).$$

结构化条件提高可控性，也降低从纯文本推断几何/时间的负担。

37.3 Scene graph 与 dynamic scene graph

静态 scene graph：

$$\mathcal{G}_t = (V_t, E_t),$$

节点为对象，边为关系。动态版本增加状态轨迹和事件：

$$\mathcal{G}_{1:T} = \{\mathcal{G}_t, \mathcal{E}_{t \to t+1}\}_{t=1}^{T-1}.$$

生成可分解：

$$p(x, \mathcal{G} \mid y) = p_\omega(\mathcal{G} \mid y) p_\theta(x \mid \mathcal{G}, y).$$

评测也可检查 renderer 是否遵循 oracle graph，与 Bernini 的 oracle plan 诊断同构。

37.4 轨迹条件与对象级控制

给对象 i 的 2D/3D trajectory：

$$\tau_i = (p_{i,1}, \dots, p_{i,T}).$$

将轨迹 rasterize 成 heatmap、flow 或 token，注入 DiT。控制成功率：

$$S_{\text{traj}} = \exp\left(-\frac{1}{T}\sum_t \|\hat{p}_{i,t} - p_{i,t}\|_2/\sigma\right).$$

必须同时检查外观和物理，因为模型可能通过把对象“瞬移”到轨迹位置满足位置误差。

37.5 3D/4D 中间表示

可选表示：

- depth + camera；
- point cloud；
- Gaussian splats；
- NeRF/dynamic NeRF；
- mesh + skeleton；
- 3D feature grid；
- 4D point/scene tokens。

分解：

$$p(x_{1:T} \mid y) = \int p(x_{1:T} \mid S_{1:T}, c_{1:T}) p(S_{1:T}, c_{1:T} \mid y) dSdc.$$

优势是相机变换和跨视角一致性更自然；代价是 3D 数据少、动态拓扑复杂、重建误差可能限制生成上限。

37.6 物理 simulator 与神经 renderer

混合模型：

$$s_{t+1} = F_{\text{sim}}(s_t, a_t; \xi),$$

$$x_t \sim p_\theta(x_t \mid s_t, \text{style}, \text{camera}).$$

其中 ξ 为质量、摩擦、弹性等参数。优点：动力学可解释、可干预；缺点：需要状态初始化和 simulator coverage。可让神经模型预测 simulator 参数：

$$\hat{\xi} = g_\omega(y, x_{\text{ref}}),$$

或学习残差：

$$s_{t+1} = F_{\text{sim}}(s_t, a_t) + \Delta_\theta(s_t, a_t).$$

37.7 物理约束损失

如果可从 latent 解码状态 $\hat{s}_t = D_s(z_t)$，可加入：

动力学残差

$$\mathcal{L}_{\text{dyn}} = \sum_t \|\hat{s}_{t+1} - F(\hat{s}_t, a_t)\|^2.$$

接触约束

$$\mathcal{L}_{\text{contact}} = \sum_{i,j,t} \text{ReLU}\left(-\text{SDF}_j(p_{i,t})\right)^2.$$

几何重投影

$$\mathcal{L}_{\text{reproj}} = \sum_t \|\Pi_{t+1}(P_t) - u_{t+1}\|_1.$$

风险是状态 decoder/geometry estimator 的误差反向塑造生成器，导致 evaluator overfitting。

37.8 约束引导采样

在采样中加入能量 $E(z, y)$：

$$\tilde{v}(z_\tau) = v_\theta(z_\tau) - \lambda(\tau)\nabla_z E(z_\tau, y).$$

例如轨迹、深度或几何一致性。优点是不必完全重训；缺点是：

- latent 到物理状态的梯度不可靠；
- 强 guidance 破坏自然分布；
- 每步 evaluator 增加成本；
- 可能产生对 evaluator 的对抗样本。

37.9 因果图与干预

结构因果模型：

$$S_i = f_i(\text{Pa}(S_i), U_i).$$

观察条件与干预不同：

$$p(Y \mid X = x) \neq p(Y \mid \text{do}(X = x)).$$

在视频生成中，文本“地面湿了”可能与“下雨”相关；但干预“洒水器打开”也能导致湿地。若模型只学相关性，反事实可能失败。

可构造成对 prompt：

- 正常：球撞击积木，积木倒下；
- 干预：球从旁边经过，不接触积木；
- 反事实：球接触但被固定，积木不倒；
- 混杂：镜头切换造成视觉运动，但对象无接触。

评价模型是否对因果变量敏感、对非因果变量不敏感。

37.10 组合泛化

训练见过：

$$(A, \text{roll}), \quad (B, \text{jump}),$$

测试：

$$(A, \text{jump}), \quad (B, \text{roll}).$$

组合泛化要求 disentangle object 与 action。可定义 held-out composition split，而不是随机 clip split。

对多属性：

$$\mathcal{C} = \mathcal{O} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}.$$

测试集选择训练中未出现的笛卡尔组合，但组件单独出现。需防止近重复和语言同义泄漏。

37.11 Curriculum：从原子规律到组合世界

建议阶段：

1. 单对象静态属性；
2. 单对象简单动力学；
3. 两对象接触；
4. 多对象链式反应；
5. 相机运动 + 动态场景；
6. 多阶段因果；
7. 反事实和 OOD 参数；
8. 长程交互与策略闭环。

难度变量包括对象数 n_o、事件数 n_e、依赖深度 d_c、遮挡率、相机自由度和参数 OOD 距离。

37.12 物理数据的来源与偏差

- **互联网视频**：丰富真实外观，但动作/状态标签弱、相机混杂；
- **游戏/模拟器**：状态和动作真值强，但 sim-to-real gap；
- **机器人数据**：真实控制强，但规模和场景有限；
- **程序化合成**：可精确控制反事实，但视觉域窄；
- **生成数据**：规模大，但可能放大教师错误。

最佳策略往往是多源混合，并让每种数据负责不同监督：

$$\mathcal{L} = \lambda_w \mathcal{L}_{\text{web}} + \lambda_s \mathcal{L}_{\text{sim}} + \lambda_r \mathcal{L}_{\text{robot}} + \lambda_c \mathcal{L}_{\text{counterfactual}}.$$

37.13 研究诊断：模型真的学到物理了吗

至少检查：

1. **参数单调性**：增大初速度，位移是否单调增加；
2. **OOD 参数**：质量、摩擦、重力超出训练范围；
3. **反事实局部性**：改变一个变量是否只影响相关结果；
4. **相机不变性**：换视角后动力学判断一致；
5. **风格不变性**：真实、动画、线稿中规律一致；
6. **长程因果**：早期作用是否保留到后续；
7. **闭环效用**：规划是否改善任务结果。

如果只在相似 prompt 上获得更高 MLLM 物理分，不能证明形成了可泛化物理模型。

37.14 结构化 Planner–Dynamics–Renderer

一个可研究的统一架构：

$$y \xrightarrow{P_\omega} (\mathcal{G}_0, \mathcal{E}, \xi, c_{1:T}),$$

$$\mathcal{G}_{t+1} = F_\eta(\mathcal{G}_t, a_t, \xi),$$

$$x_{1:T} \sim R_\theta(\mathcal{G}_{1:T}, c_{1:T}, \text{style}).$$

其中：

- Planner 解析对象、事件和物理参数；
- Dynamics 更新状态；
- Renderer 负责高保真像素。

可用 oracle component 分解误差：

- oracle plan + learned dynamics + renderer；
- oracle dynamics + learned renderer；

- predicted plan + oracle dynamics；
- end-to-end。

这比只比较最终视频更能定位创新。

数学要点

物理一致性不是一个额外的视觉 loss，而是表示、数据、条件、动力学和评价的共同问题。若状态不可观测、动作不可辨识、评价器无真值，即使加入“physics reward”也可能只学到新的视觉捷径。

练习与自检

设计一个包含 20 种对象、10 种动作和 5 种物理参数的程序化训练集。建立随机 split 与 compositional split，比较纯视频 DiT、scene-graph-conditioned DiT 和 simulator-renderer 混合模型。要求同时报告 IID 质量、组合泛化、参数单调性和反事实局部性。

第 38 章 视频奖励模型：偏好数据、多维建模与 Reward Hacking

38.1 为什么基础预训练目标不等于人类偏好

Flow Matching 或 diffusion 预训练近似数据分布：

$$\min_{\theta} \mathbb{E}\|v_\theta(z_\tau, \tau, c) - u_\tau\|^2.$$

但训练数据包含：

- caption 不准确；
- 低运动或错误运动；
- 水印、字幕、剪辑；
- 不同美学和安全标准；
- 网络分布中的常见但非期望模式。

预训练学习“数据中常见什么”，偏好优化希望学习“用户或任务更想要什么”。两者并不一致。

38.2 奖励的标量与向量形式

标量奖励：

$$r_\psi(x, y) \in \mathbb{R}.$$

多维奖励：

$$\mathbf{r}_\psi(x, y) = [r_{\text{align}}, r_{\text{visual}}, r_{\text{motion}}, r_{\text{temporal}}, r_{\text{physics}}, r_{\text{safety}}].$$

标量方便优化，但掩盖 trade-off；向量更可解释，但需要聚合、约束或多目标优化。

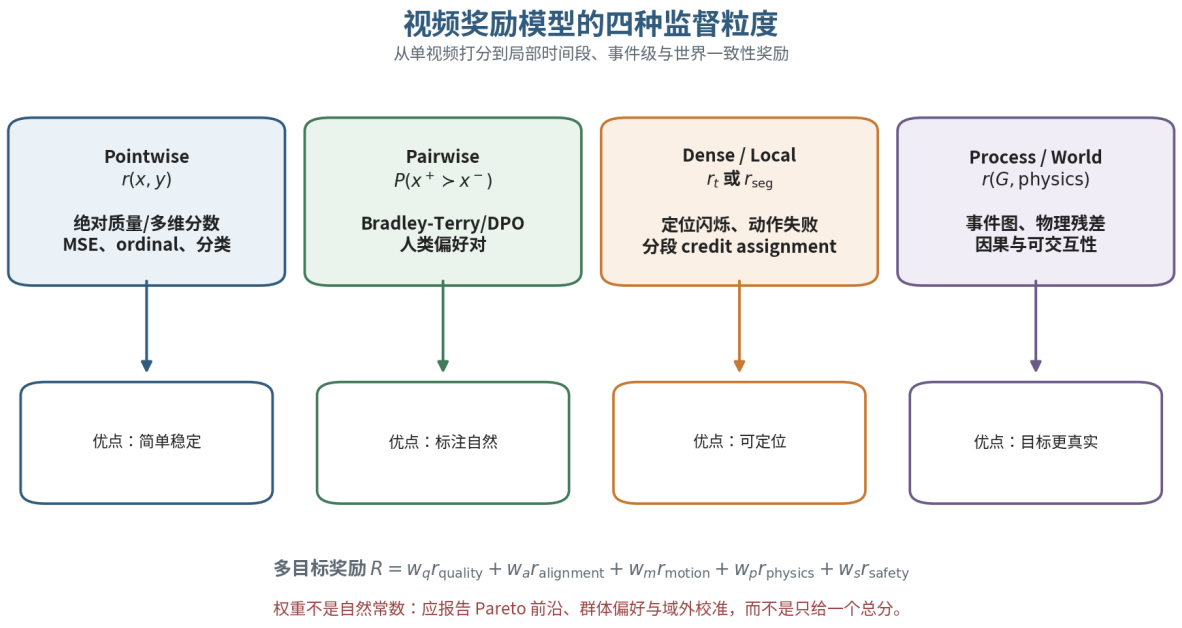

图 8: 多维视频奖励模型的训练与应用

38.3 偏好数据单位

绝对评分

$$D_{\text{score}} = \{(y_i, x_i, \mathbf{s}_i)\}.$$

成对偏好

$$D_{\text{pair}} = \{(y_i, x_i^+, x_i^-)\}.$$

排序列表

$$x_i^{(1)} \succ x_i^{(2)} \succ \cdots \succ x_i^{(K)}.$$

过程/时间局部标注

$$\{(t_{\text{start}}, t_{\text{end}}, \text{failure type}, \text{severity})\}.$$

解释与证据

包含自然语言 rationale、原子命题和 evidence frames。过程标注成本更高，却能支持 dense reward 和可解释 judge。

38.4 Bradley–Terry 奖励模型

假设潜在效用 $r_\psi(x, y)$，偏好概率：

$$P_\psi(x^+ \succ x^- \mid y) = \sigma\left(r_\psi(x^+, y) - r_\psi(x^-, y)\right).$$

损失：

$$\mathcal{L}_{\text{BT}} = -\mathbb{E} \log \sigma\left(\Delta r_\psi\right).$$

加入 margin m_i：

$$\mathcal{L}_{\text{margin}} = -\log \sigma(\Delta r_\psi - m_i).$$

若人类偏好强度不同，margin 可由 vote ratio 或评分差估计。

38.5 Listwise 与多维监督

ListMLE：

$$\mathcal{L}_{\text{list}} = -\sum_{k=1}^{K} \log \frac{\exp r(x^{(k)})}{\sum_{j=k}^{K} \exp r(x^{(j)})}.$$

多维回归：

$$\mathcal{L}_{\text{multi}} = \sum_j \lambda_j \ell_j(\hat{r}_j, s_j).$$

若维度标签尺度不同，应标准化或使用不确定性加权：

$$\mathcal{L} = \sum_j \frac{1}{2\sigma_j^2} \mathcal{L}_j + \log \sigma_j.$$

38.6 视频奖励网络架构

Frame encoder + temporal pooling

$$e_t = f_i(x_t), \qquad h = \text{Pool}(e_{1:T}), \qquad r = g(h, f_t(y)).$$

便宜，但可能忽略顺序。

Video encoder

$$h = f_v(x_{1:T}),$$

使用时空 Transformer/VideoMAE/VideoCLIP 特征，时间更敏感。

MLLM judge

输入视频、prompt、rubric，输出分数和 rationale。优点是开放问题和可解释性；缺点是成本、偏差和不稳定。

Mixture-of-Experts reward

不同 expert 负责美学、运动、文本、物理等：

$$r = \sum_k \alpha_k(x, y) r_k(x, y), \qquad \sum_k \alpha_k = 1.$$

MJ-VIDEO 类工作将复杂视频质量分解为多方面和细粒度标准，反映奖励模型正在从单标量走向多专家 judge。

38.7 VideoFeedback、VideoScore 与 VideoScore2

VideoFeedback/VideoScore 路线的核心是构建多模型、多 prompt、多维人工评分数据，再训练统一 evaluator。其研究价值包括：

- 用真实生成失败训练，而非只用自然视频；
- 显式覆盖 visual、temporal、alignment 等维度；
- 可用于模型选择、数据过滤和后训练。

VideoScore2 进一步结合更精细的数据、解释和强化学习式训练，提高 judge 的推理和泛化。阅读此类工作应关注：

1. 训练集包含哪些生成模型；
2. 测试是否包含真正未见模型家族；
3. prompt 域和视频时长；
4. 人类标注一致性上限；
5. 是否对模型名、风格和压缩敏感；
6. 与下游后训练收益是否相关。

38.8 Dense reward：时间定位比总体分数更有用

整体奖励：

$$r(x_{1:T}, y).$$

时间局部奖励：

$$r_t = r(x_{t-w:t+w}, y, a_t).$$

事件级奖励：

$$r_e = r(x_{\tau_e^s:\tau_e^e}, a_e).$$

Dense reward 可帮助：

- 只修复失败时间段；
- 给 diffusion timestep/latent region 分配 credit；
- 训练局部 DPO；
- 避免整段视频因一个小错误被同等惩罚。

但 dense labels 本身难获得，可用 MLLM 伪标注 + 人工校准。

38.9 Reward calibration

奖励差值应对应偏好概率：

$$P(x^+ \succ x^-) \approx \sigma\left(\frac{r(x^+) - r(x^-)}{T}\right).$$

温度 T 可在 validation 上拟合。还应评估：

- reliability diagram；
- Brier score；
- ECE；
- subgroup calibration；
- tie calibration。

奖励排序准确但尺度不校准，可能仍适合 reranking，却不适合直接作为 RL reward magnitude。

38.10 Reward uncertainty

使用 ensemble 或 Bayesian approximation：

$$\mu_r(x) = \frac{1}{M}\sum_m r_m(x),$$

$$\sigma_r^2(x) = \frac{1}{M-1}\sum_m (r_m(x) - \mu_r(x))^2.$$

保守奖励：

$$r_{\mathrm{LCB}} = \mu_r - \kappa\sigma_r.$$

在 OOD prompt、长视频、动画或新模型家族中，uncertainty 应上升。若不升，说明模型过度自信。

38.11 Reward hacking 的典型形式

动态度 hacking

为提高 motion score，引入无意义相机抖动。

美学 hacking

生成漂亮但静态、简单、偏离 prompt 的画面。

CLIP hacking

反复呈现 prompt 中高权重对象，忽略关系和时间。

Judge hacking

生成 evaluator 偏好的字幕、构图、风格或伪影。

物理 reward hacking

让动作太小或对象不交互，从而避免明显物理错误。

长视频 truncation hacking

在 evaluator 只采前 32 帧时，让前段优质、后段崩坏。

38.12 防 reward hacking 的策略

1. 多维奖励 + 硬约束；
2. 多 evaluator ensemble；
3. adversarial/red-team prompts；
4. hidden holdout evaluator；
5. 定期人工审计高奖励样本；
6. reward model 更新，但避免非平稳失控；
7. 对低运动、重复、字幕、异常相机单独惩罚；
8. 在优化前后做 distribution shift 检测；
9. 对 reward 与人评差异最大的样本 active learning；
10. 监控 KL 到参考模型。

38.13 多目标奖励与约束优化

线性标量化：

$$r = \sum_k w_k r_k.$$

只能找到凸 Pareto 前沿的一部分，并高度依赖权重。可使用约束：

$$\max_\theta \ \mathbb{E}[r_{\text{pref}}]$$

$$\text{s.t.} \quad \mathbb{E}[r_{\text{align}}] \geq c_1, \quad \mathbb{E}[r_{\text{safety}}] \geq c_2, \quad D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \leq \epsilon.$$

Lagrangian：

$$\mathcal{J} = \mathbb{E}[r_{\text{pref}}] + \lambda_1(\mathbb{E}[r_{\text{align}}] - c_1) + \lambda_2(\mathbb{E}[r_{\text{safety}}] - c_2) - \beta D_{\text{KL}}.$$

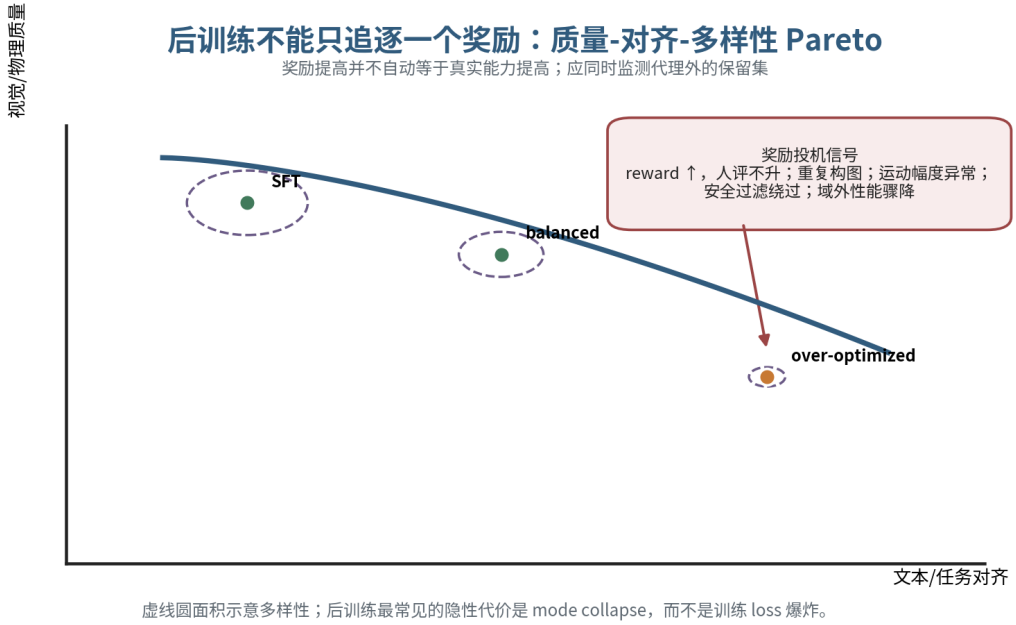

图 9: 质量、对齐、多样性与奖励投机之间的 Pareto 关系

38.14 奖励模型评测协议

In-distribution

与训练模型和 prompt 相似，测基本拟合。

Model-OOD

使用未见生成器、采样器、VAE、风格和分辨率。

Prompt-OOD

专业域、长 prompt、多语言、组合、物理和反事实。

Corruption-OOD

编码、插帧、裁剪、字幕、颜色和长程稀疏错误。

Optimization-OOD

用奖励优化后的策略生成视频，测试 reward 是否仍与人类一致。这是最关键的一类，因为策略会主动寻找 reward 漏洞。

38.15 奖励模型数据卡

至少记录：

```yaml
reward_dataset:
  prompts: 12000
  videos: 48000
  generator_families: 14
  pairwise_labels: 120000
  absolute_labels: 48000
  dimensions:
    - alignment
    - visual
    - temporal
    - motion
    - physics
  raters_per_item: 3
  tie_allowed: true
  rationales: subset_20_percent
  duration_range_sec: [2, 30]
  held_out_generator_families: 3
  held_out_prompt_domains: [robotics, animation]
```

常见误区

奖励模型的验证集若与训练集来自同一批生成模型，相关性很高并不意味着可用于优化。真正困难的测试是：当策略已经针对奖励模型优化后，奖励分数是否仍预测独立人类偏好。

第 39 章　视频后训练：SFT、Reward Gradient、DPO、蒸馏与 GRPO

39.1 后训练的统一视角

预训练得到参考模型 π_{ref}。后训练希望：

$$\max_{\pi} \mathbb{E}_{y\sim\mathcal{P},x\sim\pi(\cdot|y)}[r(x,y)] - \beta D(\pi\|\pi_{\text{ref}}).$$

第一项提高偏好，第二项防止偏离预训练分布过远。不同方法的区别在于：

- 如何表示策略概率；
- 是否需要可微奖励；
- 是否在线采样；
- credit assignment 在视频、帧、latent 还是 denoising step；
- 是否蒸馏采样速度。

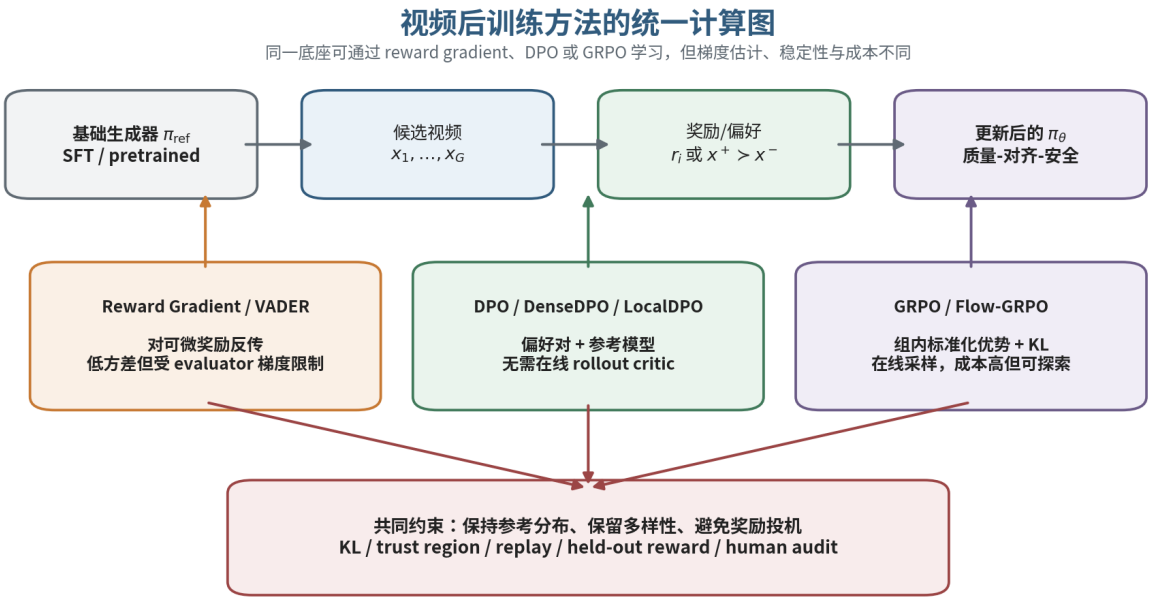

图 10: 视频后训练方法的目标与代价谱系

39.2 Supervised Fine-Tuning

高质量数据 D_{HQ} 上继续训练原始生成目标：

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(x,y)\sim D_{\text{HQ}}} \mathcal{L}_{\text{FM/diff}}(x, y).$$

优势：

- 稳定；
- 不需要奖励梯度；
- 可增强特定任务、风格或分辨率；
- 易与 LoRA/adapter 结合。

局限：

- 依赖高质量正样本；
- 无法直接利用“哪个更好”的偏好对；
- 对稀有失败覆盖不足；
- 可能过拟合小数据并遗忘基础能力。

InstructVideo 类工作通过指令化数据和人类反馈增强文本遵循，可视为视频生成 SFT/偏好数据构建的早期路线。

39.3 Reward Gradient：对采样路径反向传播

若奖励 $r_\psi(x, y)$ 可微，目标：

$$\mathcal{J}(\theta) = \mathbb{E}_\epsilon[r_\psi(G_\theta(\epsilon, y), y)].$$

梯度：

$$\nabla_\theta \mathcal{J} = \mathbb{E}\left[\frac{\partial r}{\partial x}\frac{\partial x}{\partial \theta}\right].$$

但 x 由 K 步采样得到：

$$z_{k-1} = F_\theta(z_k, k, y),$$

完整反传需存储或重算所有步，显存和计算昂贵。VADER 类方法研究如何对视频扩散模型通过 reward gradients 做对齐，并使用 truncated backprop、随机截断或参数高效微调。

Truncated backprop

只从部分采样步反传：

$$\nabla_\theta r \approx \frac{\partial r}{\partial z_0} \prod_{k=1}^{K'} \frac{\partial z_{k-1}}{\partial z_k} \frac{\partial z_{K-K'}}{\partial \theta}.$$

偏差与成本之间权衡。

39.4 Diffusion-DPO 的基本推导

语言模型 DPO 基于最优策略：

$$\pi^*(x \mid y) \propto \pi_{\text{ref}}(x \mid y) \exp(r(x, y)/\beta).$$

因此奖励差可写为

$$r(x^+, y) - r(x^-, y) = \beta \log \frac{\pi^*(x^+ \mid y)/\pi_{\text{ref}}(x^+ \mid y)}{\pi^*(x^- \mid y)/\pi_{\text{ref}}(x^- \mid y)}.$$

DPO 损失：

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \log \sigma \left(\beta \left[\log \frac{\pi_\theta(x^+ \mid y)}{\pi_{\text{ref}}(x^+ \mid y)} - \log \frac{\pi_\theta(x^- \mid y)}{\pi_{\text{ref}}(x^- \mid y)} \right] \right).$$

扩散/流模型没有容易计算的精确 log-likelihood，需使用变分界、去噪误差或路径代理。

39.5 扩散模型的偏好代理

对同一噪声时间 τ 和噪声 ϵ，定义样本去噪损失：

$$\ell_\theta(x, y; \tau, \epsilon) = \|v_\theta(z_\tau, \tau, y) - u_\tau\|^2.$$

较小损失可视为较高代理 log-probability：

$$\log \pi_\theta(x \mid y) \approx -C + \mathbb{E}_{\tau,\epsilon}[-w(\tau)\ell_\theta(x, y; \tau, \epsilon)].$$

于是偏好 logit 可构造为：

$$\Delta_\theta = -(\ell_\theta^+ - \ell_\theta^-) + (\ell_{\text{ref}}^+ - \ell_{\text{ref}}^-).$$

$$\mathcal{L}_{\text{video-DPO}} = -\log \sigma(\beta \Delta_\theta).$$

实际实现中应让正负样本共享 τ、噪声和预处理，以降低方差。

39.6 DenseDPO 与 LocalDPO

整体偏好对可能只在局部帧存在差异。DenseDPO 类方法将 reward/偏好分配到时间或空间：

$$\mathcal{L} = \sum_{t,p} \alpha_{t,p} \ell_{\text{DPO}}^{t,p}.$$

LocalDPO 进一步强调局部失败区域和时间段，减少对原本正确区域的无谓更新。关键挑战：

- 如何获得可靠 mask；
- 局部修改是否影响全局运动；
- evaluator localization 是否准确；
- 负样本是否只改变目标因素。

39.7 Discriminator-free 与 implicit preference

如果没有显式偏好模型，可利用正负数据、质量过滤或自生成对构造隐式偏好。Discriminator-free DPO 类方法避免单独训练 discriminator，但仍需某种正负定义。风险是过滤器偏差被直接写入策略。

39.8 HuViDPO 与人类视频偏好

视频偏好与图像不同：评审必须考虑运动、时间、动作和长期一致。HuViDPO 类工作强调 human video preference data 和视频 DPO。数据设计比目标函数同样重要：若负样本只是低清晰度，模型会学画质而非运动或物理。

39.9 蒸馏：T2V-Turbo 与 DOLLAR

一致性/轨迹蒸馏

教师多步采样：

$$z_{t_b} = \Phi_{\text{teacher}}(z_{t_a}, t_a \to t_b, y).$$

学生一步预测：

$$\hat{z}_{t_b} = F_\theta(z_{t_a}, t_a, t_b, y).$$

蒸馏损失：

$$\mathcal{L}_{\text{distill}} = \|\hat{z}_{t_b} - \text{sg}(z_{t_b})\|^2.$$

T2V-Turbo 将视频一致性蒸馏与奖励反馈结合，目标是在极少步数下保持质量和对齐。

DOLLAR 类多奖励蒸馏

将多个奖励和分布保持结合：

$$\mathcal{L} = \lambda_d \mathcal{L}_{\text{distill}} - \sum_k \lambda_k r_k(x, y) + \lambda_c \mathcal{L}_{\text{consistency}}.$$

蒸馏不能只看最终 VBench；还要测：

- NFE/branch-equivalent NFE；
- motion diversity；
- prompt adherence；
- teacher–student mode coverage；
- 新分辨率/时长外推；
- CFG 是否已被蒸馏。

39.10 Policy Gradient 与 GRPO

把视频生成视为策略：

$$x \sim \pi_\theta(x \mid y).$$

REINFORCE：

$$\nabla_\theta J = \mathbb{E}\left[(r(x, y) - b(y))\nabla_\theta \log \pi_\theta(x \mid y)\right].$$

扩散策略的 log-probability 可按反向转移近似分解：

$$\log \pi_\theta(\tau \mid y) = \sum_{k=1}^{K} \log p_\theta(z_{k-1} \mid z_k, y).$$

Group Relative Policy Optimization（GRPO）对同一 prompt 采样 G 个视频，组内标准化奖励：

$$A_i = \frac{r_i - \bar{r}}{s_r + \epsilon}.$$

再使用 clipped ratio：

$$\rho_i = \frac{\pi_\theta(\tau_i \mid y)}{\pi_{\text{old}}(\tau_i \mid y)},$$

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\left[\min(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i)\right] + \beta D_{\text{KL}}.$$

视频中最昂贵的是每个 prompt 需要多次完整采样和评价。

39.11 VGGRPO：latent 几何奖励

VGGRPO 将几何 foundation model 接到 diffusion latent，避免每次奖励都完整 VAE decode，并对动态场景构造：

- camera motion smoothness reward；
- geometry reprojection consistency reward。

其一般思想：

$$z \xrightarrow{L_\omega} \hat{g} \xrightarrow{r_g} r_{\text{geometry}},$$

在 latent 空间计算世界一致性奖励。优势是效率和梯度接口；风险是 latent geometry model 的偏差成为新 reward loophole。

39.12 Self-paced GRPO

固定 prompt 难度会导致：容易样本 reward 饱和，困难样本全部失败、组内方差太小。Self-paced 方法根据学习进度调整 prompt/难度采样：

$$p_t(y) \propto p_0(y) \exp\left(\lambda\, u_t(y)\right),$$

u_t 可由组内 reward variance、成功率或学习增益定义。理想课程聚焦“当前可学但未掌握”的样本。

39.13 Diffusion-DRF 与 dense reward feedback

Dense Reward Feedback 路线将 reward 分配到 denoising steps、帧或区域，从而改善 credit assignment。抽象：

$$J = \sum_{k=1}^{K} \gamma_k r_k(z_k, y),$$

而非只在最终 x 上给 reward。需要防止中间 latent reward 与最终视频质量不一致。

39.14 On-policy 与 off-policy

Off-policy preference

使用固定 D_{pair}，如 DPO。稳定、便宜，但数据很快落后于当前策略。

On-policy RL

当前策略生成样本，再评分。适应策略分布，但昂贵且更易 reward hacking。

混合

$$D_t = \alpha D_{\text{offline}} + (1-\alpha) D_{\text{onpolicy},t}.$$

定期加入人工审计和 hard negatives。

39.15 KL、模型坍缩与多样性

偏好优化常导致：

- 风格收缩；
- 低运动；
- prompt 模板化；
- 多样性下降；
- reward 过拟合。

监控参考 KL 的代理：

$$\widehat{D}_{\text{KL}} = \mathbb{E}\left[\log \pi_\theta(\tau \mid y) - \log \pi_{\text{ref}}(\tau \mid y)\right].$$

以及同 prompt 多 seed 的 diversity：

$$D_{\text{seed}} = \frac{2}{K(K-1)} \sum_{i<j} d(\phi(x_i), \phi(x_j)).$$

39.16 参数高效后训练

对 14B 视频 DiT，常只训练：

- LoRA on self-attn/cross-attn；
- timestep modulation；
- adapter/control branch；
- expert subset；
- output head；
- planner 或 reward bridge。

LoRA：

$$W' = W + \frac{\alpha}{r} BA, \quad A \in \mathbb{R}^{r \times d_{\text{in}}}, B \in \mathbb{R}^{d_{\text{out}} \times r}.$$

但后训练是否需要更新 self-attention 取决于目标：文本遵循可偏 cross-attention，运动/物理通常涉及 self-attention 和 temporal representation。

39.17 一个安全的后训练阶段表

1. **SFT warm start**：高质量、多样正样本；
2. **offline DPO**：人类偏好对；
3. **reward model OOD 校准**；
4. **small on-policy stage**：保守 KL；
5. **independent evaluator + human audit**；
6. **distillation**：减少 NFE；
7. **distilled model 再审计**；
8. **红队和发布门控**。

不要同时改变奖励、采样器、数据和架构后只报告最终提升，否则无法定位原因。

39.18 后训练报告清单

```yaml
post_training:
  base_checkpoint: sha256:...
  method: video_dpo
  trainable_modules: [self_attn_lora, cross_attn_lora]
  preference_pairs: 120000
  generators_in_data: 10
  beta: 0.1
  timestep_sampling: logit_normal
  shared_noise_for_pair: true
  reference_model: frozen
  optimizer: adamw
  steps: 10000
  global_batch_pairs: 128

monitoring:
  proxy_kl: true
  reward_train: true
  reward_holdout: true
  independent_judge: true
  human_audit_every_steps: 1000
  diversity: true
  motion_floor: true
```

核心结论

后训练并不是“让 reward 变高”，而是在参考模型约束下改善经过独立校准的人类/任务偏好，同时保持多样性、物理、文本和安全。优化器直接面对的 reward 与论文用来证明成功的 evaluator 最好不是同一个模型。

练习与自检

为一个 Wan2.1-1.3B 底座设计三组可比实验：SFT、Video-DPO、reward-gradient LoRA。固定训练视频数量、GPU-hours 和推理配置；对每种方法报告训练 reward、独立 reward、人评、seed diversity、motion、KL proxy 和 OOD prompt。解释哪一组证据能证明不是 reward hacking。

第 40 章　测试时扩展：采样、搜索、批评、修复与蒸馏

40.1 为什么 test-time compute 重新重要

训练后模型固定，但仍可通过增加推理计算改善结果：

$$\text{quality} = f(\text{model}, \text{NFE}, \text{candidate count}, \text{critic}, \text{repair iterations}).$$

传统视频生成主要增加 sampler steps；更一般的 test-time scaling 包括：

- 同 prompt 多候选；
- reward reranking；
- 分层计划搜索；
- MLLM critique；
- 局部重采样和编辑；
- 长视频状态验证与回滚；
- 多 agent 审核。

因此比较模型时必须同时报告 test-time budget。

40.2 Best-of-N

生成

$$x_i \sim \pi_\theta(x \mid y), \quad i = 1, \dots, N,$$

选择

$$x^* = \arg\max_i r_\psi(x_i, y).$$

若单样本 reward CDF 为 $F(r)$，最大值 CDF：

$$F_{\max}(r) = F(r)^N.$$

随着 N 增大，期望最大值提高，但边际收益递减。若 reward 有噪声：

$$\hat{r}_i = r_i + \epsilon_i,$$

大 N 会增加选择到“reward noise 极值”的风险，即 winner’s curse。

正确报告

- N；
- 每候选 NFE；
- 是否共享 prompt rewrite/初始噪声；
- ranking evaluator；
- 最终独立 evaluator；
- 总 GPU-seconds；
- 人工挑选是否存在。

40.3 Rejection sampling

接受概率：

$$P(\text{accept} \mid x) = \mathbb{I}[r(x) \geq \tau]$$

或 soft acceptance：

$$P(\text{accept} \mid x) \propto \exp(r(x)/T).$$

阈值过高会导致：

- 成本爆炸；
- 多样性下降；
- reward hacking；
- 困难 prompt 几乎无样本被接受。

应按 prompt 难度自适应阈值，或在预算内选择 Pareto 候选。

40.4 多目标候选选择

候选分数向量

$$\mathbf{r}_i = (r_i^{\text{align}}, r_i^{\text{visual}}, r_i^{\text{motion}}, r_i^{\text{physics}}).$$

不要简单固定线性权重。可先删除被 Pareto 支配的候选：若存在 j 满足

$$r_j^k \geq r_i^k \ \forall k$$

且至少一维严格大，则 i 被支配。再根据应用约束选：

$$r^{\text{motion}} \geq c_m, \quad r^{\text{safety}} \geq c_s.$$

40.5 Prompt/plan 搜索

把扩写后的 prompt 或 semantic plan 当作搜索变量 p：

$$p^* = \arg\max_p \mathbb{E}_{x \sim G(\cdot|p)} r(x, y),$$

约束

$$\text{SemEq}(p, y) \geq \tau.$$

候选 plan 可以包含：

- shot sequence；
- camera；
- object states；
- action timeline；
- negative constraints；
- reference allocation。

需要防止 planner 为迎合 reward 改变用户意图。

40.6 Tree search over storyboard

长视频可对故事状态做搜索。节点是部分 storyboard $S_{1:k}$，扩展候选下一 shot：

$$S_{k+1}^{(j)} \sim P_\omega(\cdot \mid S_{\leq k}, y).$$

价值：

$$V(S_{\leq k}) = q_{\text{coherence}} + q_{\text{coverage}} + \widehat{V}_{\text{future}}.$$

Beam search 保留 top-B。但 planner score 未必预测 renderer 成功率，最好用低成本 proxy render 对候选做验证。

40.7 MLLM critique–revise

迭代：

$$x^{(0)} = G(y),$$

$$c^{(k)} = Q_\omega(x^{(k)}, y),$$

$$x^{(k+1)} = E_\theta(x^{(k)}, c^{(k)}, y).$$

critique 应输出：

- 失败类型；
- 时间区间；
- 受影响对象；
- 保留区域；
- 修复指令；
- 置信度。

例如：

```
{
  "failure": "the cup becomes empty before pouring ends",
  "interval_sec": [2.1, 3.8],
  "objects": ["cup", "water"],
  "preserve": ["person_identity", "camera", "background"],
  "repair": "regenerate the liquid and cup state only",
  "confidence": 0.92
}
```

40.8 局部时空重采样

给 mask $M \in \{0,1\}^{T\times H\times W}$，保留未遮罩区域：

$$z_\tau = M \odot z_\tau^{\text{new}} + (1-M) \odot z_\tau^{\text{orig}}.$$

在每个去噪步对未编辑区域重新加噪并注入，可实现 video inpainting。关键是：

- mask 时间边界要平滑；
- 被编辑区域需继承 motion/identity；
- 物理变化可能需要扩展到因果相关区域；
- 只改后果不改原因会产生新不一致。

40.9 VideoRepair

VideoRepair 类方法先检测低质量区域或时间段，再局部再生成。它体现“生成–理解–修复”闭环。评价不能只比较修复区，还需测：

$$\Delta Q_{\text{target}}, \quad \Delta Q_{\text{preserve}}, \quad \Delta Q_{\text{global}}.$$

理想：目标改善，保留区变化小，全局不降低。

40.10 ImagerySearch 与生成空间搜索

ImagerySearch 类研究将视觉生成视为可迭代搜索：提出候选、观察、批评、更新。对于复杂物理或构图，搜索可能比单次 prompt 更有效。

抽象为：

$$h_{k+1} = U(h_k, x_k, r_k, c_k),$$

$$q_{k+1} = P(h_{k+1}, y),$$

$$x_{k+1} = G(q_{k+1}).$$

需要评估搜索是否真正改进，而非通过更多样本和 cherry-pick。

40.11 长视频中的验证与回滚

每生成一个 chunk，检查：

- 身份 drift；
- 状态表冲突；
- 重复；
- 物理 failure；
- 叙事 coverage；
- 安全。

若置信度低，可：

1. 重采当前 chunk；
2. 回滚到上一个稳定状态；
3. 增强 reference；
4. 修改局部 plan；
5. 重新生成多个后续候选。

这类似数据库事务：只把通过验证的状态 commit 到 memory。

40.12 Test-time scaling curve

对预算 C 画质量：

$$Q(C).$$

成本包括：

$$C = C_{\text{generate}} + C_{\text{judge}} + C_{\text{repair}} + C_{\text{planner}}.$$

比较应使用：

- 固定总 GPU-seconds；
- 固定 latency；
- 固定候选数；
- 或报告完整 Pareto curve。

只把单次生成的 latency 与经过 16 候选 reranking 的质量放在一起是不公平的。

40.13 Search over reward 的过拟合

选择最大 noisy reward 时，独立真实质量的期望可能先升后降。对候选数 N，应画：

$$\hat{r}_{\text{selection}}(N), \quad r_{\text{independent}}(N), \quad h_{\text{human}}(N).$$

若 selection reward 继续升，而 independent/human 停滞或下降，说明 search 正在 exploit evaluator。

40.14 推理时审计日志

```
{
  "request_id": "...",
  "base_prompt": "...",
  "rewritten_prompt": "...",
```

```
  "candidate_count": 8,
  "candidate_seeds": [1,2,3,4,5,6,7,8],
  "generator_nfe_each": 20,
  "ranking_model": "reward_v3_sha256...",
  "ranking_scores": [0.4,0.7,0.6,0.8,0.5,0.3,0.65,0.55],
  "selected_candidate": 3,
  "critic_calls": 1,
  "repair_calls": 1,
  "repair_intervals": [[2.0, 3.5]],
  "total_gpu_seconds": 184.2,
  "safety_decision": "allow",
  "provenance_manifest": "..."
}
```

日志对复现、成本和安全追踪同样重要。

核心结论

测试时扩展把“模型能力”变成“模型 + 预算 + 搜索器 + evaluator”的系统能力。任何质量对比都应同时报告推理预算和选择机制；否则较高分可能只是更多抽样或更强人工筛选的结果。

练习与自检

对固定底座比较：40-step 单样本、10-step Best-of-4、10-step Best-of-4 + 一轮局部修复。固定总 branch-equivalent NFE，报告质量、独立人评、reward gap、延迟与多样性。分析哪种方案在相同预算下更优。

40.15 少步蒸馏与质量—成本 Pareto

本节承接前述测试时搜索，讨论另一条降低推理成本的路线：将多步生成轨迹蒸馏为少步模型。

40.15.1 Progressive distillation

逐步将教师的两步映射蒸馏为学生一步：

$$G^{(1)}_{\theta_s}(z_t, t) \approx G^{(2)}_{\theta_T}(z_t, t).$$

反复减半 NFE。训练稳定，但多轮成本高。

40.15.2 Consistency model

不同时间点沿同一 probability flow trajectory 映射到同一数据：

$$f_\theta(z_t, t) \approx f_\theta(z_s, s) \approx z_0.$$

可实现一步/少步，但视频高维和长序列下训练更难。

40.15.3 Distribution matching distillation

学生生成分布匹配教师分布，通过 score/reward/discriminator 差异更新。CausVid 将其用于因果视频学生；T2V-Turbo 等方法将少步蒸馏与 reward feedback 结合。

40.15.4 Rectified flow/ReFlow

让运输轨迹更直，降低 ODE 求解步数：

$$\min_\theta \mathbb{E}\|v_\theta(z_t, t) - (z_1 - z_0)\|^2,$$

并通过模型配对重新生成路径。视频中需防止少步带来的动态僵硬和高频丢失。

40.16 对齐与蒸馏的顺序

三种策略：

1. `align → distill`：先对齐高步模型，再蒸馏；可能丢失偏好；
2. `distill → align`：先少步，再后训练；少步模型探索能力弱；
3. joint alignment-distillation：共同优化，可能梯度冲突但更直接。

设损失：

$$\mathcal{L} = \lambda_d \mathcal{L}_{\text{distill}} + \lambda_r \mathcal{L}_{\text{reward}} + \lambda_k \mathcal{L}_{\text{KL}}.$$

梯度冲突可测 cosine：

$$\cos(g_d, g_r) = \frac{g_d^\top g_r}{\|g_d\|\|g_r\|}.$$

若长期为负，可采用 PCGrad、动态权重、参数分支或共享同构表征。Reward Lightning 的动机正是减少不同表征空间造成的冲突。

40.17 少步模型的评测陷阱

- 固定 steps 不等于固定 NFE/分支；
- CFG 蒸馏后 guidance 口径不同；
- FPS 可能不含 text encoder/VAE/I/O；
- batch 1 与高 batch throughput 混淆；
- 低分辨率上实时不代表目标分辨率；
- 首次编译和 warm-up 被排除；
- 只报告最佳短 prompt；
- 少步模型可能降低多样性；
- reward evaluator 可能偏爱过锐化。

推荐报告：

$$\text{Pareto point} = (\text{quality}, \text{alignment}, \text{motion}, \text{diversity}, \text{latency}, \text{VRAM}, \text{energy}).$$

40.18 动态计算分配

对简单 prompt 用少步，对复杂 prompt 用更多计算。难度预测器：

$$\hat{d}(q) = D_\psi(q),$$

策略：

$$K(q) = K_{\min} + (K_{\max} - K_{\min})\sigma(\hat{d}(q)).$$

也可在采样中根据 residual/confidence 提前停止。必须校准：困难 prompt 被错误判简单时会系统性失败。

40.19 质量-成本实验

至少比较：

配置	NFE	CFG 分支	候选数	Planner	Repair	总延迟
Base-50	50	2	1	否	否	…
Student-4	4	1	1	否	否	…
Student-4 BoN4	4	1	4	否	否	…
Student-4 + plan	4	1	1	是	否	…
Student-4 + repair	4	1	1–2	是	是	…

同等 latency 下比较最终效用，才能回答“更聪明的测试时计算”是否优于“更强的单次模型”。

40.20 本章自检

练习与自检

1. 为什么 Best-of-100 可能比 Best-of-10 更容易 reward hacking?2. 设计一个局部重生成的 collateral damage 指标。3. 比较 align-then-distill、distill-then-align 和 joint 的优缺点。4. 如何测量两个损失的梯度冲突？5. 给定固定延迟预算，如何在 NFE、候选数、planner 和 judge 之间分配?

第 41 章　安全、来源证明、版权、隐私与发布治理

41.1 安全不是一个 classifier

视频生成风险分布在全生命周期：

$$\text{data} \to \text{pretraining} \to \text{post-training} \to \text{prompt} \to \text{generation} \to \text{distribution} \to \text{reuse}.$$

单一输入过滤器或输出 NSFW classifier 无法覆盖：

- 身份冒用和 deepfake；
- 非自愿亲密内容；
- 欺诈和虚假证据；
- 未成年人风险；
- 暴力、自残和违法指导；
- 隐私与训练数据记忆；
- 版权和风格模仿；
- 偏见与代表性伤害；
- 误导性政治/公共信息；
- 物理 AI 中的错误模拟和危险决策；
- 水印移除和来源伪造。

因此需要 defense-in-depth。

图 11: 可信视频生成的纵深防御体系

41.2 风险模型：能力、可达性、意图与影响

可用半定量风险：

$$R = P(\text{hazard}) \times P(\text{exposure} \mid \text{hazard}) \times I(\text{impact}).$$

进一步分解：

$$P(\text{hazard}) = f(C, A, U, M),$$

- C：模型能力；
- A：攻击者可访问性；
- U：用户意图和使用上下文；
- M：缓解措施。

同一模型在离线研究 checkpoint、带身份控制的公开 API 和无日志匿名下载中的风险不同。

41.3 NIST Generative AI Profile 的工程映射

NIST AI 600-1 将生成式 AI 风险管理组织在 Govern、Map、Measure、Manage 四类活动中。映射到视频生成：

Govern

- 明确责任人和发布门槛；
- 维护模型卡、数据卡和事件响应；
- 供应链与第三方 evaluator 管理；
- 红队和风险接受机制。

Map

- 定义目标用户、任务和不可接受用途；
- 识别受影响群体；
- 描述数据来源、身份与版权风险；
- 分析部署场景和滥用路径。

Measure

- 安全 benchmark；
- memorization/privacy 测试；
- deepfake/身份相似度；
- 水印鲁棒性；
- 偏见和语言覆盖；
- 误报/漏报；
- OOD 与不确定性。

Manage

- 模型访问控制；
- 速率限制；
- 内容过滤；
- 来源凭证；
- 举报、撤下和补救；
- 持续监控与更新。

框架提供管理结构，不代替具体法律判断。

41.4 数据治理

来源和许可

对每个数据源记录：

- 来源 URL/平台/供应商；

- 获取时间；
- 许可和使用条件；
- 是否允许机器学习训练；
- 是否包含人物、生物特征、未成年人；
- 地理与语言范围；
- 删除请求和版本。

可追溯 manifest

```
{
  "asset_id": "sha256:...",
  "source_type": "licensed_archive",
  "source_uri_hash": "...",
  "license_id": "...",
  "consent_scope": ["research_training"],
  "faces_detected": 2,
  "minors_flag": "unknown",
  "caption_model": "...",
  "filter_version": "...",
  "duplicate_group": "dg_1042",
  "deletion_status": "active"
}
```

删除与重训

理想系统能从 asset 追踪到：

$$\text{asset} \to \text{clip} \to \text{shard} \to \text{checkpoint lineage}.$$

大模型完全机器遗忘仍困难，但资产级索引、训练快照和后续版本删除是最低要求。

41.5 训练数据记忆与隐私

记忆风险可通过：

- prompt-based extraction；
- nearest-neighbor retrieval；
- membership inference；
- canary exposure；
- identity similarity；
- rare sequence 复现。

若插入 canary c，exposure：

$$\text{exposure}(c) = \log_2 |\mathcal{R}| - \log_2 \text{rank}(c),$$

概念来自语言模型记忆测试，可改造为视频特征检索。视频中的“近重复”需在多层表示上判断：

- frame perceptual hash；
- face embedding；
- video embedding；
- motion pattern；
- shot sequence。

41.6 身份与 deepfake 风险

身份生成能力包括：

- text-only 生成真实人物；
- reference image/video 驱动；
- face swap；
- voice/audio 联合合成；
- 动作与场景伪造。

缓解可分：

1. **输入授权**：reference ownership/consent；
2. **身份策略**：公共人物、私人个体、未成年人分级；
3. **输出检测**：face identity similarity、敏感场景；
4. **来源标记**：watermark + content credentials；
5. **访问控制**：高风险身份功能限制；
6. **追踪与响应**：日志、举报、撤下。

人脸相似度阈值会有种族、年龄、姿态和画质差异，必须做 subgroup calibration。

41.7 水印：可检测信号而非真实性证明

视频水印通常将消息 m 嵌入：

$$\tilde{x} = W_\eta(x, m),$$

检测器：

$$\widehat{m} = D_\omega(\tilde{x}').$$

训练考虑攻击 $\mathcal{A}$：

$$\mathcal{L} = \mathbb{E}_{x,m,a\sim\mathcal{A}}\left[\ell(D_\omega(a(W_\eta(x,m))), m)\right] + \lambda d(x, \tilde{x}).$$

攻击包括：

- H.264/H.265/AV1 压缩；
- resize/crop；
- 帧率改变；
- 插帧/删帧；
- 颜色和噪声；
- 局部编辑；
- screen recording；
- 再生成；
- watermark removal model。

检测指标

- bit accuracy；
- TPR at fixed FPR；
- video-level detection；
- 定位能力；
- 鲁棒性–感知质量 Pareto；
- false attribution 风险。

VideoMark、VIDSTAMP、VideoShield 等研究分别探索视频水印、时空鲁棒性和 diffusion latent 中的标记。

水印不能证明什么

- 无水印不等于真实；
- 有水印不等于内容事实正确；
- 检测失败不等于未生成；
- 水印可被移除、复制或伪造；
- 不同供应商不可天然互操作。

41.8 C2PA 与 Content Credentials

C2PA 的核心是对内容来源和编辑历史建立签名 manifest。概念上：

$$\text{asset} + \text{claims} + \text{actions} + \text{ingredients} \xrightarrow{\text{digital signature}} \text{verifiable manifest}.$$

典型声明：

- 由某生成系统创建；
- 使用的模型或工具；
- 发生的裁剪、调色、生成式编辑；
- 引用的输入资产；
- 签名实体和时间。

Content Credentials 与水印互补：

机制	优势	局限
签名 manifest	信息丰富、可验证、记录编辑链	元数据可能被剥离
隐形水印	与像素绑定、可在元数据丢失时检测	容量小、攻击下可能失效
可见标签	用户直观	易裁剪、影响体验
平台日志	可追责、上下文丰富	跨平台不可见

理想系统将 manifest、watermark 和日志绑定为同一 asset ID。

41.9 EU AI Act Article 50 的透明度背景

欧盟《AI Act》第 50 条包含对某些 AI 生成或操纵内容的透明度义务，包括 deepfake 场景的披露要求，并对艺术、讽刺等场景规定相应表达方式。该法规的一般适用日期是 **2026 年 8 月 2 日**；本教材编写日期为 2026 年 7 月 12 日，因此这一日期尚未来临。实际适用需结合角色、用途、地区、具体内容和后续实施指南判断。

工程上可提前准备：

- 机器可读来源标记；
- 面向用户的清晰披露；
- 生成/编辑操作日志；
- deepfake 场景识别；
- 艺术/创作场景的适当标识；
- 供应商与部署者责任边界。

本节是技术合规概览，不构成法律意见。

41.10 版权与人类创作

美国版权局关于生成式 AI 可版权性的报告强调人类作者贡献的重要性：仅由 AI 自动产生的材料与包含足够人类创作选择、安排或修改的作品，法律分析不同。

技术系统应保留人类创作过程证据：

- prompt 和 storyboard；
- 多轮选择与编辑；
- mask/trajectory/镜头设计；
- 人工后期；
- 版本历史；
- 素材许可。

但“记录了 prompt”不自动证明满足任何法域的可版权标准；需按具体事实和法律判断。

41.11 训练版权风险的技术面

需要区分：

- 训练数据获取和许可；
- 模型是否记忆/近似复现；
- 输出与特定作品的实质相似；
- 商标、肖像和不正当竞争；
- 风格模仿与作者归属。

技术缓解：

1. licensed/opt-in 数据池；
2. source filtering；
3. near-duplicate removal；
4. 输出相似度检索；
5. 特定角色/标志过滤；
6. opt-out 与删除机制；
7. 生成 provenance；
8. 高相似输出人工复核。

相似度模型只是筛查工具，不是法律裁判。

41.12 安全分类器的级联

输入：

$$q_{\text{in}} = C_{\text{text}}(y, r),$$

其中 r 是 reference。生成后：

$$\mathbf{q}_{\text{out}} = C_{\text{video}}(x, y).$$

策略：

$$\text{decision} = \Pi(q_{\text{in}}, \mathbf{q}_{\text{out}}, \text{user tier}, \text{jurisdiction}, \text{context}).$$

分类器应输出类别和置信度，而非只给 allow/block。对高风险、低置信样本可转人工。

41.13 多模态规避攻击

攻击可能把风险分散到：

- 文本隐喻；
- reference image；
- 首尾帧；
- 控制轨迹；
- 多轮编辑；
- 低分辨率或局部区域；
- 长视频后半段。

因此安全评测必须覆盖完整输入组合和多轮工作流。只扫描 prompt 文本不足。

41.14 长视频安全

长视频会产生新问题：

- 敏感内容只在少量帧出现；
- 角色身份在后段漂移成真实人物；
- 安全故事逐步转为危险场景；
- evaluator 只看均匀抽帧而漏检；
- 多镜头上下文使单镜头含义改变。

采用层级扫描：

1. 稀疏全局抽帧；
2. shot-level scan；
3. 高风险 shot 密集扫描；
4. audio/text/OCR 联合；
5. 跨镜头语义；
6. 人工复核。

41.15 世界模型的安全特殊性

Physical AI 世界模型的错误可能影响真实动作。需额外关注：

- 危险动作后果被低估；
- OOD 场景过度自信；
- 规划器利用模型漏洞；
- 模型模拟与真实系统不一致；
- 合成数据造成策略偏差；
- 安全约束在长 rollout 中遗忘。

安全规划可使用 pessimistic value：

$$V_{\text{safe}} = \mathbb{E}[V] - \kappa\sqrt{\text{Var}(V)}.$$

并设置 hard constraints 与真实系统 shield。

41.16 红队矩阵

轴	示例
内容	sexual、violence、self-harm、fraud、hate
身份	celebrity、private person、minor、employee
模态	text、image、video、audio、trajectory
语言	多语言、谐音、编码、隐喻
工作流	single-shot、edit、multi-turn、long video
攻击	jailbreak、classifier evasion、watermark removal
分发	截图、重编码、拼接、平台转发
风险结果	生成成功率、漏检率、可追踪性、响应时间

应区分 model capability 与 deployed system capability，红队两者都要测。

41.17 发布分级

权重开放

优点是研究透明和可复现；风险是安全控制无法强制、微调可移除限制。

托管 API

可日志、更新、速率限制和撤回；但集中化且无法完全防止输出再利用。

分层访问

按能力或用途开放：

- 基础 T2V；
- 真实人物 reference；
- 长视频；
- 去水印/编辑；
- 动作条件世界模型。

高风险能力可要求身份验证、用途说明、研究协议或更严格审计。

41.18 模型卡中的安全段落

```yaml
safety:
  intended_uses:
    - research_on_video_generation
    - licensed_creative_content
  excluded_uses:
    - nonconsensual_identity_impersonation
    - deceptive_evidence
```

```yaml
high_risk_capabilities:
  - reference_identity_transfer
  - long_video_generation
input_filters:
  version: safety_text_image_v4
output_filters:
  version: video_safety_v3
provenance:
  c2pa_manifest: true
  invisible_watermark: true
logging:
  retention_days: 30
red_team_report: redteam_2026q2.pdf
known_limitations:
  - multilingual_evasion
  - sparse_long_video_failures
  - watermark_degradation_after_heavy_editing
```

41.19 事件响应

当发现严重输出：

1. 保存证据和请求上下文；
2. 限制相关能力或账户；
3. 复现并确定攻击路径；
4. 评估影响范围；
5. 更新过滤器/策略/模型；
6. 通知受影响方和必要机构；
7. 提供撤下和申诉；
8. 发布透明度说明；
9. 将案例加入回归测试。

安全不是发布前一次性测试，而是持续控制循环。

41.20 安全指标

对风险类别 c：

$$\mathrm{ASR}_c = \frac{\#\text{成功产生违规输出}}{\#\text{攻击请求}}.$$

过滤器：

$$\mathrm{TPR}, \mathrm{FPR}, \mathrm{Precision}, \mathrm{Recall}.$$

还应报告：

- subgroup rates；
- multilingual ASR；
- multi-turn ASR；
- long-video sparse failure recall；
- watermark TPR at fixed FPR；
- provenance retention after transformations；
- 举报到处置时间；
- 安全机制的用户效用成本。

常见误区

“带水印”不等于“安全”，“通过内容分类器”不等于“合法”，“公开模型卡”不等于“风险已管理”。来源、内容、身份、隐私、版权和真实世界后果是不同问题，需要不同控制和证据。

练习与自检

为一个支持参考图 I2V、最长 60 秒、可生成人物的视频 API 建立 threat model。列出资产、攻击者、攻击路径、风险类别、预防、检测、响应和剩余风险。然后为 text-only、reference identity 和 long-video 三种能力设计分级访问策略。

第 42 章 研究选题地图：从可复现问题到博士级贡献

42.1 什么构成视频生成研究贡献

一个有说服力的研究项目通常同时满足：

$$\text{重要问题} + \text{明确假设} + \text{可识别干预} + \text{可信测量} + \text{外推证据}.$$

“加入模块后 VBench 高 0.5” 通常不够。需要说明：

1. 模块针对哪类失败；
2. 为什么现有表示/目标无法解决；
3. 改变了模型设计向量的哪一维；
4. 如何排除数据、采样和计算差异；
5. 哪个 oracle 或反事实支持机制解释；
6. 是否在新 prompt、长时域或新模型上复现。

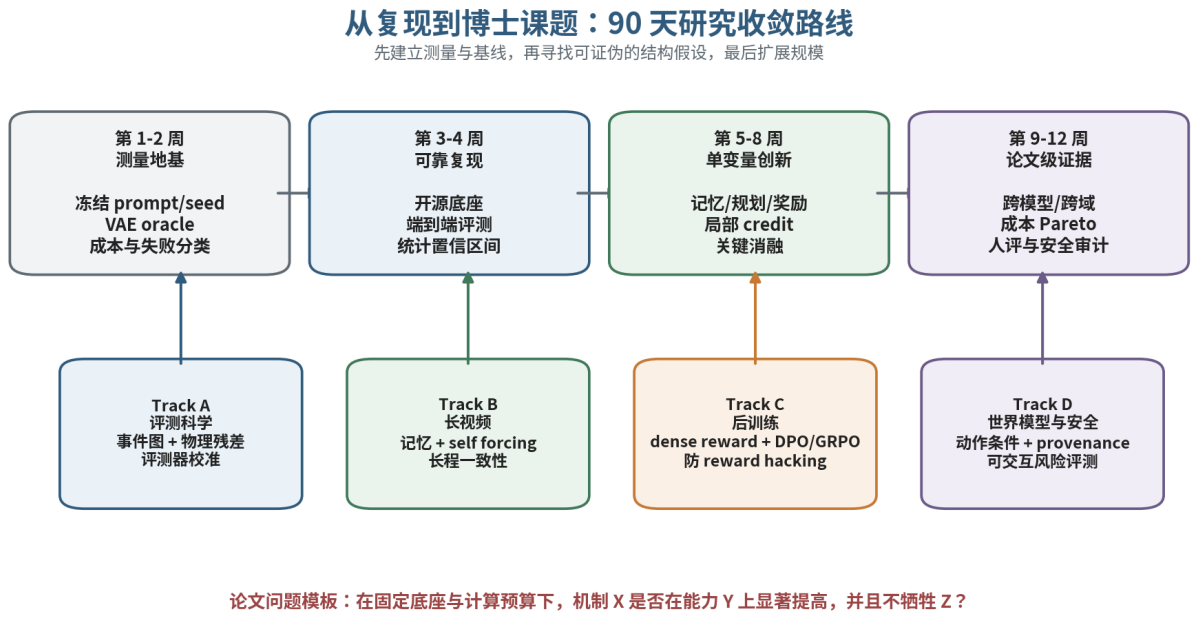

图 12: 视频生成研究选题地图

42.2 研究假设模板

不要写：

> “我们提出一个更好的 temporal module。”

应写：

> “当生成长度超过训练 horizon 时，短窗口 Video DiT 的身份漂移主要来自生成历史进入条件后造成的分布偏移，而不是 attention 感受野不足。若训练时按受控比例注入模型生成前缀，并使用置信度门控的身份记忆，则跨 4 倍训练 horizon 的身份漂移应下降，同时不增加运动冻结。”

可验证预测：

- real-history 条件下各方法接近；
- generated-history 条件下 baseline 明显下降；
- self-history 训练降低漂移；
- 只加更大 attention window 不完全解决；
- 过强 memory 会提高 freeze；
- 在未见身份和场景上仍成立。

42.3 项目一：Evaluator Stress Lab

研究问题

现有视频 evaluator 对哪些受控退化敏感，哪些存在盲区？

方法

建立变换集合

$$\mathcal{T} = \{\text{flicker}, \text{shuffle}, \text{freeze}, \text{id swap}, \text{trajectory error}, \text{physics reversal}, \text{sparse long failure}\}.$$

每种退化有强度 λ 和局部时间 mask。对 FVD、JEDi、VBench、VideoScore、MLLM judge 做：

- pairwise accuracy；
- monotonicity；
- localization；
- invariance；
- human correlation；
- OOD model transfer。

贡献形式

- benchmark + metric audit；
- 新的 evaluator ensemble/calibration；
- “指标失效条件”比新总分更有价值。

计算预算

低：不训练生成器；需要视频变换、评测模型和小规模人评。

42.4 项目二：VAE-aware Evaluation Ceiling

假设

不同生成模型的部分排名差异来自 VAE reconstruction ceiling，而非 DiT。

实验

对多个 VAE：

$$x \to E_i(x) \to D_i(E_i(x)).$$

测：

- FVD/JEDi；
- OCR、face identity、flow、physics state；
- 细粒度失败；
- 不同压缩率和 channel。

再将多个 DiT 的分数做 ceiling-normalized：

$$Q_{\text{norm}} = \frac{Q_{\text{gen}} - Q_{\text{noise}}}{Q_{\text{oracle}} - Q_{\text{noise}}}.$$

风险

指标方向和线性归一化未必合理，应报告原始分与归一化分。

42.5 项目三：Generated-history Curriculum for Long Video

假设

暴露偏差是长视频失稳的重要原因；混合真实和自生成历史可改善稳定性。

训练分布：

$$h_t \sim \alpha_t p_{\text{real}}(h) + (1-\alpha_t) p_\theta(h).$$

逐步减小 α_t。比较：

- teacher forcing；
- random corruption；
- self-history；
- confidence-filtered self-history；
- self-history + memory reset。

评价到 1x、2x、4x、8x horizon，测 drift 曲线和恢复率。

42.6 项目四：Typed Memory for Long Video

问题

单一 keyframe/latent memory 混合身份、几何、语义和近期运动，易污染。

方法

$$m_t = (m_t^{\text{id}}, m_t^{\text{geom}}, m_t^{\text{event}}, m_t^{\text{recent}}).$$

不同写入规则：

$$m_{t+1}^k = g_k(m_t^k, o_t; \text{conf}_k).$$

不同读取 gate：

$$h' = h + \sum_k \alpha_k(h,t)\,\text{Attn}(h, m^k).$$

关键消融

- untyped concat；
- typed but no confidence；
- typed + confidence；
- oracle memory；
- memory corruption。

42.7 项目五：Planner–Renderer Interface Benchmark

动机

Bernini/分层模型收益可能受 planner、renderer 或接口限制。

三组计划

$$\text{no-plan}, \quad \text{predicted-plan}, \quad \text{oracle-plan}.$$

再构造 corrupted plan：

- identity swap；
- event reversal；
- position perturbation；
- semantic token dropout；
- segment-ID collision。

测 renderer sensitivity：

$$S_{\text{plan}} = \frac{\partial Q_{\text{output}}}{\partial Q_{\text{plan}}}.$$

贡献

建立连续视觉语义条件的接口测试，而非只比较最终模型。

42.8 项目六：Latent Geometry Reward without Hacking

动机

VGGRPO 表明 latent geometry reward 可减少 VAE decode 成本，但 reward model 可能被策略 exploit。

方法

训练 latent geometry model L_ω，同时保留独立 RGB geometry evaluator E。优化用 L_ω，验证用 E 和人评。

$$\max_\theta r_L(z) \quad \text{s.t.} \quad r_E(D(z)) \geq c, \quad D_{\text{KL}} \leq \epsilon.$$

加入 adversarial holdout scene 和 dynamic object。

判据

若训练 reward 升、独立 geometry 和 closed-loop 也升，才支持世界一致性改善。

42.9 项目七：Physics Counterfactual Generator

数据

从 simulator 生成四元组：

$$(s_0, a, \xi, x),$$

并构造只改变一个因素的反事实：

$$(s_0, a', \xi, x'), \quad (s_0, a, \xi', x'').$$

模型

结构化 Planner–Dynamics–Renderer，或视频 DiT + state adapter。

指标

- intervention locality；
- parameter monotonicity；
- trajectory error；
- visual quality；
- compositional OOD；
- action identification。

博士级扩展

从程序化域迁移到互联网视频，学习可解释 latent physical variables。

42.10 项目八：Reward Model Generalization under Optimization

核心问题

reward model 在静态 held-out 上准确，是否在被策略优化后仍准确？

协议

1. 训练 reward R_1；
2. 用 R_1 对生成器做不同强度优化；
3. 用独立 reward R_2、人类和受控退化测；
4. 画 optimizer steps 对 score 的曲线。

定义 reward overoptimization gap：

$$G(t) = R_1(\pi_t) - H(\pi_t),$$

其中 H 是人类归一化分数。研究 early stopping、ensemble、uncertainty penalty 和 active learning。

42.11 项目九：Long-video MLLM Judge with Evidence

问题

长视频 judge 常漏掉稀疏故障，且无法说明依据。

方法

层级检索：

1. shot/segment summarization；
2. query-conditioned candidate interval retrieval；
3. dense verification；
4. cross-event reasoning；
5. evidence timestamp 输出。

损失：

$$\mathcal{L} = \mathcal{L}_{\text{answer}} + \lambda_1 \mathcal{L}_{\text{interval}} + \lambda_2 \mathcal{L}_{\text{calibration}}.$$

数据

使用合成 needle distortions + 人类验证，再迁移到真实长生成视频。

42.12 项目十：World-model Validity beyond Pixels

假设

像素指标对策略排序的预测能力有限；状态和动作一致性更重要。

实验

在可控游戏/机器人环境中训练多个世界模型。对每个模型：

- FVD/LPIPS；
- state prediction；
- action response；
- calibration；
- simulated policy return；
- real policy return。

比较：

$$\text{Corr}(\text{pixel metric}, J_{\text{real}})$$

与

$$\text{Corr}(\text{state/action metric}, J_{\text{real}}).$$

若后者显著更强，可推动世界模型评价范式改变。

42.13 项目十一：Watermark–Provenance Joint Robustness

问题

manifest 易被剥离，水印信息有限；如何联合恢复来源链?

方法

水印编码 asset ID，C2PA manifest 存详细声明，服务端 transparency log 保存映射。测试：

- 压缩、裁剪、插帧；
- 局部编辑；
- screen recording；
- 再生成；
- 拼接多个来源；
- 恶意 manifest 替换。

指标：

$$\text{TPR@FPR}, \quad \text{attribution accuracy}, \quad \text{tamper localization}, \quad \text{false accusation rate}.$$

42.14 项目十二：Quality–Cost–Safety Pareto Scaling

动机

模型论文往往只比较质量，而真实部署同时受成本和风险约束。

定义向量：

$$\mathbf{u}(M) = (Q, -C, -L, S, D),$$

- Q：质量；
- C：GPU 成本；
- L：延迟；
- S：安全；
- D：多样性。

系统比较：模型大小、VAE 压缩、NFE、Best-of-N、蒸馏、量化和 safety layers。输出 Pareto frontier 和应用条件，而非单一冠军。

42.15 计算预算分级

Tier 0：单卡/免费资源

- evaluator meta-evaluation；
- VAE oracle；
- 已生成视频分析；
- 小模型 reward calibration；
- benchmark/data 工具。

Tier 1：1–8 张 80GB GPU

- 1B–5B LoRA；
- adapter/control；
- 小规模 DPO；
- long-video inference method；
- reward model fine-tuning。

Tier 2：16–64 张 GPU

- 5B–14B 参数高效后训练；
- on-policy video RL；-较大 planner；
- 多域长视频训练。

Tier 3：基础预训练

需要大规模数据、集群和工程团队。博士研究不必以从零训练 14B 为起点；清晰的问题、强诊断和可复现证据往往比扩大参数更有贡献。

42.16 一个 12 周研究启动计划

第 1–2 周：复现测量

- 固定一个开源底座；
- 建 manifest；
- 复现官方推理；
- 计算形状、成本和 VAE oracle；
- 跑多维 benchmark。

第 3–4 周：失败数据集

- 选择一个核心失败；
- 构建 200–1000 个控制 prompt；
- 人工检查 evaluator；
- 建立 baseline 和置信区间。

第 5–7 周：最小方法

- 只改一个接口；
- 低成本 LoRA/adapter；
- 记录训练和推理成本；
- 完成关键 oracle 消融。

第 8–9 周：外推

- 新模型底座或新数据域；
- 2x–4x 时长；
- OOD 组合；
- evaluator stress test。

第 10–11 周：人评和统计

- 预注册主终点；
- pilot power；
- 盲化配对；
- bootstrap/BT；
- 失败 taxonomy。

第 12 周：论文故事

- 问题证据；
- 机制假设；
- 方法；
- 单变量消融；
- 外推；
- 局限和风险。

42.17 论文实验表的推荐结构

主表

报告多维质量 + 成本：

Model	Align	Visual	Motion	Temporal	Physics	Human Pref	NFE	GPU-s

机制表

Variant	Oracle condition	Generated history	Typed memory	Key outcome

长程曲线

$1\times, 2\times, 4\times, 8\times$ horizon，而非只有最终长度。

失败表

按身份、几何、运动、物理、文本、重复分类。

安全与外推表

新模型、新域、多语言、稀疏失败和 reward-optimized distribution。

42.18 研究中的负结果

有价值的负结果包括：

- 新指标在人类 meta-evaluation 上不优于简单 baseline；
- 更强 identity memory 导致 motion freeze；
- physics reward 提升 judge 分但不提升状态真值；
- 4-step 蒸馏提高速度却严重收缩多样性；
- MLLM planner 的 oracle plan 有效，但 predicted plan 无收益；
- 长视频局部质量不降，但全局事件重复上升。

负结果若有清晰控制和机制解释，可以直接指导后续研究。

42.19 研究伦理与可发布性

在立项时就判断：

- 数据是否有权使用；
- 人评是否需要伦理审批/同意；
- 是否涉及真实人物或敏感内容；
- checkpoint 开放是否增加高风险能力；
- benchmark 是否包含有害样本，怎样安全存储；
- 日志是否包含个人数据；
- 水印和来源是否保留。

方法性能与发布方式可以分离：论文可公开算法和低风险评测，而对高风险权重采用分层访问。

附录 A 统计与人类评测公式速查

A.1 均值差与配对效应量

对 prompt-level 差值 d_q：

$$\bar{d} = \frac{1}{Q}\sum_q d_q, \qquad s_d^2 = \frac{1}{Q-1}\sum_q (d_q - \bar{d})^2,$$

$$d_z = \frac{\bar{d}}{s_d}.$$

A.2 Bootstrap percentile CI

1. 从 Q 个 prompt 有放回采样；
2. 计算 $\bar{d}^{(b)}$；
3. 重复 B 次；
4. 取 $\alpha/2$ 和 $1-\alpha/2$ 分位数。

对于小样本/偏斜分布可使用 BCa bootstrap。

A.3 Spearman 与 Kendall

Spearman：对 rank 做 Pearson。

$$\rho_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$$

（无 ties 的简化式）。

Kendall：

$$\tau = \frac{N_{\text{concordant}} - N_{\text{discordant}}}{\binom{n}{2}}.$$

A.4 Cohen’s κ

$$\kappa = \frac{p_o - p_e}{1 - p_e}.$$

类别极不平衡时 κ 可能反直觉，应同时报告 confusion matrix。

A.5 Krippendorff α

$$\alpha = 1 - \frac{D_o}{D_e}.$$

支持多个评审、缺失值和不同测量尺度；距离函数应与标签类型匹配。

A.6 Brier 与 ECE

$$\text{BS} = \frac{1}{n}\sum_i (p_i - y_i)^2.$$

$$\text{ECE} = \sum_b \frac{|I_b|}{n} |\operatorname{acc}(I_b) - \operatorname{conf}(I_b)|.$$

A.7 Bradley-Terry

$$P(i \succ j) = \sigma(\theta_i - \theta_j).$$

带 prompt 随机效应：

$$P(i \succ j \mid q) = \sigma(\theta_i - \theta_j + u_{i,q} - u_{j,q}).$$

A.8 Holm 修正

将 $p_{(1)} \le \cdots \le p_{(m)}$ 排序，依次比较：

$$p_{(k)} \le \frac{\alpha}{m-k+1}.$$

A.9 Benjamini-Hochberg

找到最大 k 满足：

$$p_{(k)} \le \frac{k}{m}\alpha,$$

拒绝 $1, \dots, k$，控制 FDR。

A.10 层级 bootstrap 建议

- 推广到新 prompt：先采样 prompt；
- 同 prompt 多 seed：在 prompt 内再采样 seed；
- 人评多 rater：可在 pair 内采样 rater；
- 模型比较保持配对，不要分别独立采样两模型。

附录 B 视频生成评测基准卡片

Benchmark	年份	重点	适合回答	主要限制
FVD	2018/2019	分布距离	整体真实/生成特征差异	特征、样本量、内容偏差
VBench	2023/2024	多维技术质量	画质、时间、基本对齐 profile	部分维度趋于饱和
VBench++	2024	T2V/I2V 与可信性扩展	跨任务综合比较	总分依赖协议
EvalCrafter	2023/2024	多指标开放域评测	模型能力矩阵	自动指标相关不等于效度
T2V-CompBench	2024	组合语义	属性/关系/动作绑定	VLM judge 与 prompt 范围
ChronoMagic-Bench	2024	time-lapse	长过程与状态变化	特定时间推移域
VideoPhy	2024	物理常识	prompt 与物理联合	规模和评测器边界
VideoPhy-2	2025	200 类动作物理	动作中心物理与规则	开放视频不可完全量化
PhyCoBench	2025	多类物理原则	物理专项诊断	相机/尺度不可识别
T2VPhysBench	2025	多物理规律	分规律比较	judge 与任务定义敏感
T2VTextBench	2025	视频文字	OCR、稳定与动态文字	OCR 对风格/遮挡敏感
VBench-2.0	2025	内在真实性	人体、控制、创造、物理、常识	自动 judge 仍需校准
SafeSora	2024	安全与偏好数据	安全 reward/对齐	类别与时代覆盖
SafeGen-Bench	2026	条件视频安全	多模态组合风险	新基准，需持续验证
WorldReasonBench	2026	世界推理	多步状态、物理/因果	与闭环控制仍有距离
CultureScore	2026	文化表征	身份/情境/行为	文化定义与标注群体

使用卡片时，应进一步锁定论文版本、代码 commit、prompt 和 evaluator。

附录 C 后训练公式速查

C.1 Reward-weighted SFT

$$\mathcal{L} = -\mathbb{E}[w(R)\ell_\theta(x,q)].$$

C.2 DPO

$$\mathcal{L}_{\mathrm{DPO}} = -\log\sigma\left(\beta[(\ell_\theta^+ - \ell_{\mathrm{ref}}^+) - (\ell_\theta^- - \ell_{\mathrm{ref}}^-)]\right).$$

C.3 Dense/local DPO

$$\mathcal{L} = \sum_k \lambda_k \mathcal{L}_{\mathrm{DPO}}(\ell_{\theta,k}^+, \ell_{\theta,k}^-).$$

C.4 Reward gradient

$$\nabla_\theta \mathbb{E}[R(G_\theta(\epsilon, q))] = \mathbb{E}\left[\nabla_x R \frac{\partial G_\theta}{\partial \theta}\right].$$

C.5 GRPO

$$A_i = \frac{r_i - \bar{r}}{s_r + \epsilon},$$

$$\mathcal{L} = -\mathbb{E}\min(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon_c, 1+\epsilon_c)A_i) + \beta \text{KL}.$$

C.6 Constrained objective

$$\mathcal{L} = -J_{\text{main}} + \sum_j \lambda_j (c_j - J_j), \qquad \lambda_j \leftarrow [\lambda_j + \eta(c_j - J_j)]_+.$$

C.7 多样性保留

$$\mathcal{L}_{\text{div}} = -\mathbb{E}_{i \neq j} d(F(x_i), F(x_j)).$$

注意不要鼓励语义错误造成的差异。

附录 D 长视频与世界模型分类表

系统	是否动作条件	是否因果流式	长期记忆	是否规划验证	主要目标
普通 T2V DiT	否	通常否	无/文本	否	短视频生成
Chunk T2V	否	可选	overlap/ embedding	否	更长开放环视频
Story Planner-Renderer	高层事件	可选	角色/道具/故事状态	局部	多镜头叙事
Causal streaming video	可选	是	KV/cache/窗口	否	实时连续生成
Future predictor	无显式动作	可选	历史 state	否	视频预测
GameNGen 类	是	是	帧/动作历史	可用于交互	特定游戏引擎
Genie/通用交互世界	是	是	世界状态	潜在/显式	开放交互环境
Robot world model	连续控制	是/离线 rollout	belief state	是	控制与规划

附录 E 安全与来源证明模板

E.1 风险登记表

Risk ID	场景	严重度	可能性	检测	缓解	Owner	Residual

E.2 红队样本记录

```
{
  "prompt_id": "...",
  "attack_family": "multimodal_identity_bypass",
  "language": "zh",
```

```
  "reference_hash": "...",
  "policy_version": "...",
  "model_version": "...",
  "result": "allowed|refused|partial",
  "harm_severity": 4,
  "detection_stage": "output_segment_3",
  "watermark_valid": true,
  "c2pa_valid": true,
  "reviewer": "..."
}
```

E.3 Incident response

- 发现时间；
- 影响版本/用户；
- 风险类别与严重度；
- 临时 containment；
- 证据和日志保存；
- 模型/策略修复；
- 用户/监管/合作方通知；
- 回溯测试；
- benchmark 与数据更新；
- 复盘 owner 和 deadline。

附录 F　核心术语表

术语	定义
construct validity	指标是否真正测量目标能力
reliability	重复测量的一致性
calibration	预测置信度与真实正确率是否匹配
paired bootstrap	保持同一 prompt 配对结构的重采样 CI
event graph	实体、状态、事件以及时间/因果边组成的图
object permanence	对象遮挡、切镜或时间延迟后仍保持身份和状态
intrinsic faithfulness	超越表面质量的物理、常识、人体和组合真实性
exposure bias	训练用真实历史、推理用模型历史导致的分布差异
self forcing	在模型自身 rollout 历史上训练/蒸馏以减少 mismatch
semantic memory	用视觉/语言 embedding 保存跨段高层信息
structured state	显式角色、道具、地点和事件状态
action-conditioned world model	根据历史和动作预测未来观测/状态的模型
counterfactual branch	同一历史下因不同动作产生的不同未来
model exploitation	规划器利用世界模型错误获得虚假高回报
pointwise reward	对单条视频给绝对分数
pairwise reward	对两个候选预测偏好概率
dense reward	对帧/时间段提供局部反馈
process reward	对事件、因果或物理过程评分
reward hacking	模型提高代理 reward 而未提高真实目标
likelihood displacement	DPO 中 winner 绝对似然也可能下降
GRPO	以同 prompt 组内相对优势进行策略优化
homologous preference distillation	在共享表征中联合偏好与蒸馏的思路
hard binding	用加密 hash 将 C2PA manifest 绑定到资产
soft binding	用指纹/不可见水印恢复 manifest 关联
Content Credentials	C2PA manifest 的用户友好称呼
false refusal	安全系统错误拒绝无害请求
attack success rate	对抗请求绕过安全措施的比例

附录 G　随附工具

本册源文件包包含以下可运行脚本：

```
python tools/paired_bootstrap_eval.py --help
python tools/bradley_terry_fit.py --help
python tools/event_graph_eval.py --help
python tools/reward_model_audit.py --help
python tools/long_video_state_audit.py --help
```

其用途分别为：

- 从逐 prompt/seed CSV 计算配对差值、bootstrap CI 和 win rate；
- 从 pairwise 人评拟合 Bradley-Terry 分数；
- 对目标/预测事件图计算节点、时间边和因果边指标；
- 审计 reward 与人评的相关、校准、subgroup 和潜在投机；
- 对长视频逐 chunk 状态记录计算召回、身份漂移和退化斜率。

工具是研究脚手架，不替代对数据 schema 和统计假设的检查。

参考文献与公开资料

以下按主题整理；年份为首次公开或主要版本年份。2026 年条目多为预印本，应关注后续修订。

评测与指标

1. Salimans et al. *Improved Techniques for Training GANs*. NeurIPS, 2016.（Inception Score）
2. Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. NeurIPS, 2017.（FID）
3. Unterthiner et al. *Towards Accurate Generative Models of Video: A New Metric and Challenges*. 2018/2019.（FVD）
4. Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. ICML, 2021.（CLIP）
5. Huang et al. *VBench: Comprehensive Benchmark Suite for Video Generative Models*. CVPR, 2024. arXiv:2311.17982.
6. Huang et al. *VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models*. 2024. arXiv:2411.13503.
7. Zheng et al. *VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness*. 2025. arXiv:2503.21755.
8. Liu et al. *EvalCrafter: Benchmarking and Evaluating Large Video Generation Models*. CVPR, 2024. arXiv:2310.11440.
9. Sun et al. *T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-Video Generation*. 2024. arXiv:2407.14505.
10. Yuan et al. *ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Time-lapse Video Generation*. 2024. arXiv:2406.18522.
11. Bansal et al. *VideoPhy: Evaluating Physical Commonsense for Video Generation*. 2024. arXiv:2406.03520.
12. Bansal et al. *VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation*. 2025. arXiv:2503.06800.
13. PhyCoBench authors. *PhyCoBench: A Physical Commonsense Benchmark for Text-to-Video Generation*. 2025. arXiv:2502.05503.
14. T2VPhysBench authors. *T2VPhysBench: A First-Principles Benchmark for Physical Consistency in Text-to-Video Generation*. 2025. arXiv:2505.00337.
15. T2VTextBench authors. *T2VTextBench: A Benchmark for Text Rendering in Text-to-Video Models*. 2025. arXiv:2505.04946.
16. Ge et al. *T2VScore: Discriminative Metrics for Text-to-Video Generation*. 2024. arXiv:2401.07781.
17. VideoScore authors. *VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation*. 2024. arXiv:2406.15252.
18. Xu et al. *VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation*. 2024. arXiv:2412.21059.
19. MJ-VIDEO authors. *MJ-VIDEO: Fine-Grained Video Generation Evaluation and Reward Modeling*. 2025. arXiv:2502.01719.
20. *FVD is Not Enough: Content Bias in Video Generation Metrics*. 2024. arXiv:2404.12391.
21. *Beyond FVD: Enhanced Distribution Metrics for Video Generation / JEDi*. 2024. arXiv:2410.05203.
22. *STREAM: A Streaming Evaluation Framework for Video Generation*. 2024. arXiv:2403.09669.
23. *WorldReasonBench: Human-Aligned Stress Testing of Video Generators as Future World-State Predictors*. 2026. arXiv:2605.10434.
24. *WorldBench: Disambiguating Physics for Diagnostic Evaluation of World Models*. 2026. arXiv:2601.21282.
25. *MIND: Benchmarking Memory Consistency and Action Control in World Models*. 2026. arXiv:2602.08025.
26. *Omni-WorldBench: Towards a Comprehensive Interaction-Centric Evaluation for World Models*. 2026. arXiv:2603.22212.
27. *CULTURESCORE: Evaluating Cultural Faithfulness in Video Generation Models*. 2026. arXiv:2606.07311.

长视频、故事与流式生成

28. *StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text*. 2024. arXiv:2403.14773.
29. Yin et al. *From Slow Bidirectional to Fast Autoregressive Video Diffusion Models*. CVPR, 2025. arXiv:2412.07772.（CausVid）
30. *Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion*. 2025. arXiv:2506.08009; *Self-Forcing++: Towards Minute-Scale High-Quality Video Generation*. 2025. arXiv:2510.02283.
31. *Autoregressive Adversarial Post-Training for Real-Time Interactive Video Generation*. 2025. arXiv:2506.09350.（AAPT）

32. *LongLive: Real-Time Interactive Long Video Generation*. 2025. arXiv:2509.22622.
33. *Rolling Forcing: Autoregressive Long Video Diffusion in Real Time*. 2025. arXiv:2509.25161.
34. *Causal Forcing: Autoregressive Diffusion Distillation Done Right for High-Quality Real-Time Interactive Video Generation*. 2026. arXiv:2602.02214.
35. *Ca2-VDM: Efficient Autoregressive Video Diffusion Model with Causal Generation and Cache Sharing*. 2024. arXiv:2411.16375.
36. *VideoGen-of-Thought: Step-by-Step Video Generation*. 2025. arXiv:2503.15138.
37. *OneStory: Coherent Multi-Shot Video Generation with Adaptive Memory*. 2025. arXiv:2512.07802.
38. *StoryMem: Multi-shot Long Video Storytelling with Memory*. 2025. arXiv:2512.19539.
39. *StoryBench: A Benchmark for Story Visualization*. 2023. arXiv:2308.11606.

世界模型

40. Bruce et al. *Genie: Generative Interactive Environments*. 2024. arXiv:2402.15391.
41. Valevski et al. *Diffusion Models Are Real-Time Game Engines*. ICLR, 2025. arXiv:2408.14837.（GameNGen）
42. *GameGen-X: Interactive Open-world Game Video Generation*. 2024. arXiv:2411.00769.
43. NVIDIA. *Cosmos World Foundation Model Platform for Physical AI*. 2025. arXiv:2501.03575.
44. NVIDIA. *Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control*. 2025. arXiv:2503.14492.
45. NVIDIA. *World Simulation with Video Foundation Models for Physical AI*. 2025. arXiv:2511.00062.（Cosmos-Predict2.5）；NVIDIA. *Cosmos 3: Omnimodal World Models for Physical AI*. 2026. arXiv:2606.02800.
46. Google DeepMind. *Genie 3: A New Frontier for World Models*. Official technical blog, 2025-08-05.
47. *VRAG: Learning World Models for Interactive Video Generation*. 2025. arXiv:2505.21996.
48. *Can Test-Time Scaling Improve World Foundation Model?* 2025. arXiv:2503.24320.（SWIFT）
49. Ha and Schmidhuber. *World Models*. 2018.
50. Hafner et al. *Dreamer / Mastering Diverse Domains through World Models*. 2019–2023.

奖励模型与后训练

51. *InstructVideo: Instructing Video Diffusion Models with Human Feedback*. 2023. arXiv:2312.12490.
52. Prabhudesai et al. *Video Diffusion Alignment via Reward Gradients*. 2024. arXiv:2407.08737.（VADER）
53. Liu et al. *VideoDPO: Omni-Preference Alignment for Video Diffusion Generation*. 2024. arXiv:2412.14167.
54. *LiFT: Leveraging Human Feedback for Text-to-Video Model Alignment*. 2024. arXiv:2412.04814.
55. Wu et al. *DenseDPO: Fine-Grained Temporal Preference Optimization for Video Diffusion Models*. NeurIPS, 2025. arXiv:2506.03517.
56. *Mind the Generative Details: Direct Localized Detail Preference Optimization for Video Diffusion Models*. 2026. arXiv:2601.04068.（LocalDPO）
57. *Discriminator-Free Direct Preference Optimization for Video Diffusion*. 2025. arXiv:2504.08542.
58. Xue et al. *DanceGRPO: Unleashing GRPO on Visual Generation*. 2025. arXiv:2505.07818.
59. *Flow-GRPO: Training Flow Matching Models via Online Reinforcement Learning*. 2025. arXiv:2505.05470.
60. *A Systematic Post-Train Framework for Video Generation*. 2026. arXiv:2604.25427.
61. *Flash-GRPO: Efficient Alignment for Video Diffusion via One-Step Policy Optimization*. 2026. arXiv:2605.15980.
62. *VGGRPO: Towards World-Consistent Video Generation with 4D Latent Reward*. 2026. arXiv:2603.26599.
63. *Diverse Video Generation with Determinantal Point Process-Guided Policy Optimization*. 2025. arXiv:2511.20647.（DPP-GRPO）
64. Cheng et al. *Reward Lightning: Fast Video Generation via Homologous Preference Distillation*. 2026. arXiv:2607.03960.
65. *T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback*. 2024. arXiv:2405.18750.
66. Wallace et al. *Diffusion Model Alignment Using Direct Preference Optimization*. CVPR, 2024.
67. Black et al. *Training Diffusion Models with Reinforcement Learning*. 2023.
68. Rafailov et al. *Direct Preference Optimization: Your Language Model Is Secretly a Reward Model*. NeurIPS, 2023.
69. Schulman et al. *Proximal Policy Optimization Algorithms*. 2017.
70. *Beyond Reward Margin: Rethinking and Resolving Likelihood Displacement in Diffusion Models via Video Generation*. 2025. arXiv:2511.19049.

蒸馏与快速采样

71. Salimans and Ho. *Progressive Distillation for Fast Sampling of Diffusion Models*. ICLR, 2022.

72. Song et al. *Consistency Models*. ICML, 2023.
73. Sauer et al. *Adversarial Diffusion Distillation*. ECCV, 2024.
74. Liu et al. *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow*. ICLR, 2023.
75. Lipman et al. *Flow Matching for Generative Modeling*. ICLR, 2023.

安全、来源与治理

76. *SafeSora: Towards Safety Alignment of Text-to-Video Generation*. 2024. arXiv:2406.14477.
77. *SafeGen-Bench: Benchmarking Safety in Image-Conditioned Text-to-Video Generation*. 2026. arXiv:2606.01481.
78. *Memorization in Video Diffusion Models*. 2024. arXiv:2410.21669.
79. *RobustSora: Benchmarking Detection of AI-Generated Videos under Real-World Transformations*. 2025. arXiv:2512.10248.
80. *VIDSTAMP: A Robust Watermarking Method for Generated Video*. 2025. arXiv:2505.01406.
81. *VideoShield: Protecting Video Generation with Robust Watermarks*. 2025. arXiv:2501.14195.
82. *VideoMark: Watermarking for Video Generative Models*. 2025. arXiv:2504.16359.
83. Coalition for Content Provenance and Authenticity. *C2PA Technical Specification 2.4*. 2025/2026 current specification.
84. NIST. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*. 2024.
85. European Union. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*, especially Article 50.
86. European Commission. *AI Act - Regulatory Framework and Implementation Timeline*. Current official guidance.

全书结语：从生成概率到可验证世界

三册的统一主线可概括为：

$$\underbrace{p_\theta(\mathbf{x} \mid q, \mathbf{c})}_{\text{生成}} \longrightarrow \underbrace{\mathcal{G}, \mathbf{S}, \delta_{\text{phys}}}_{\text{结构化理解与测量}} \longrightarrow \underbrace{R, \text{DPO/GRPO}, \text{planner}}_{\text{对齐与控制}} \longrightarrow \underbrace{\text{provenance, safety, governance}}_{\text{可信部署}}.$$

上册回答“现代视频生成的数学和网络是什么”；中册回答“主流模型怎样实现、训练和复现”；下册回答“怎样证明它学到了正确能力、怎样扩展到长期和交互、怎样用偏好后训练、以及怎样控制风险”。

真正成熟的视频生成系统不是单个 DiT，而是：

$$\begin{aligned}\text{Mature Video Generation System} = {} & \text{Data Governance} + \text{Video Representation} + \text{Generator} \\ & + \text{Planner/World State} + \text{Reward/Post-training} \\ & + \text{Evaluation Science} + \text{Safety/Provenance} + \text{Operations}.\end{aligned}$$

未来模型会继续变大、压缩更强、采样更快，但最困难的问题将越来越集中在：

- 世界状态能否在长时间和动作分支中保持；
- 物理和因果是否只是视觉共现；
- 奖励能否忠实反映人类和任务价值；
- 自动评测是否可靠且不被优化；
- 用户是否知道内容的来源、修改与责任；
- 系统在失败时能否被发现、限制和修复。

对研究者而言，最有价值的能力不是追逐每周的新模型名，而是建立统一符号、明确构念、设计负对照、量化不确定性、审计副作用，并把每一项改进写成可复算的证据。做到这一点，就能够从“会调用视频模型”走向“能够研究下一代视频与世界模型”。