研究报告

Ground ID: video-55_20260416144423 **来源:** YouTube 视频"视频生成·下：模型和训练【论文精读·55】" **日期:** 2026-04-16

1. 执行摘要

视频生成已迅速成为应用 AI 领域最具竞争力的方向之一，OpenAI（SORA）、Meta（Movie Gen）和腾讯/辉览（HunyuanVideo）等主要实验室正在向一个共享的三阶段架构收敛：基于潜空间自编码器的视频压缩、基于 Transformer 骨干网络的 Flow Matching 训练生成、以及像素空间解码。本报告综合了 15 篇已开放论文的证据，对一场涵盖这三个旗舰系统及其更广泛背景（包括世界模拟器和 AGI 路线图）的学术讲座中的观点和观察进行了分析。

文献为讲座中的若干核心观点提供了强有力的实证支持。最优传输 Flow Matching 持续优于基于扩散的训练（Movie Gen 的消融研究表 8 证实）。基于 LLaMA3 的 Transformer 骨干网络在质量和文本对齐方面优于基于 DiT 的设计，尽管扩展定律表明架构差异在大规模下会减小。商业级视频生成质量现在可以约 20 万美元实现（Open-Sora 2.0），相比早期估计降低了 5–10 倍。

然而，文献也揭示了重要的未解决张力。2.5D 与 3D VAE 的权衡仍然经验上未解决——2.5D TAE 的效率提升是真实的，但伴随着边缘的感知权衡。自回归模型（VideoAR）已证明可以用减少 10 倍的推理步骤达到与扩散模型相当的质量，挑战了 Flow Matching 是确定路径的隐含假设。视频生成与世界模拟（NVIDIA Cosmos、Google DeepMind Genie 2）的融合正在加速，但生成环境的物理准确性从根本上仍局限于视觉相关性而非显式物理。这些未解决的问题定义了任何新视频生成项目的关键决策点。

来源材料中的一个[未验证声明——讲座来源]断言认为，世界模拟器是 AGI 的重要基础设施，提供用于合成数据的自我博弈环境和真值反馈。这一断言未在已发表文献中独立确立，应被视为讲座的推测性定位，而非经过验证的研究结论。

2. 问题设定与来源背景

来源材料是一场系统性讲解视频生成模型架构和训练方法的单人学术讲座（论文精读形式）。演讲者深入介绍了三个代表性工作——SORA（OpenAI）、Movie Gen（Meta）和 HunyuanVideo（腾讯/辉览）——涵盖以下技术组件：

- **视频分块与词元化**：原始视频（如 1080P、30fps）约比图像块词元大 1200 倍，远超 Transformer 的上下文限制。讲座解释说，视频压缩网络先将视频压缩到低维潜空间，然后通过 3D 卷积层（核 1×2×2）将潜表示分解为 2×2 空间块，展平为 1D 序列供 Transformer 处理。

- **视频压缩网络（VCN）设计**：讲座涵盖两种对比方法——Movie Gen 的时序自编码器（TAE，一种从 2D 图像 VAE 扩展而来的 2.5D 设计）和 HunyuanVideo 的 3D 因果 VAE（CausalConv3D）。2.5D 方法在 T/H/W 各维度实现 8×压缩（总计 512×），通道数降至 16；3D 方法在 T÷4、H÷8、W÷8 下实现相同通道压缩。讲座指出，选择 2.5D 而非 3D 是出于计算效率考虑，尽管存在边缘的感知质量权衡。
- **骨干架构**：Movie Gen 的 LLaMA3 解码器 Transformer（30B 参数，16 秒 16fps 视频的 73K 词元上下文）是参考设计，具有三个关键改进：用于多编码器文本条件的交叉注意力、用于时间步/条件注入的 AdaLN 块、以及由 Flow Matching 的非自回归特性启用的全双向自注意力。
- **训练目标**：Flow Matching，特别是最优传输（OT）Flow Matching，被呈现为主导训练目标。讲座提供了具体公式 $Xt = (1-(1-\sigma\min)t)\cdot x0 + t\cdot x1$，并解释了 OT 路径（直线插值）为何优于标准扩散的弯曲路径。
- **更广泛背景**：[未验证声明——讲座来源]讲座讨论了世界模拟器作为 AGI 重要基础设施的作用，提供用于合成数据的自我博弈环境和真值反馈。讲座概述了 AI 路线图：感知 AI → 生成式 AI（当前浪潮）→ 代理 AI（2025 年）→ 物理 AI（机器人，2-3 年内）。

讲座中识别的关键约束和开放问题包括：VAE/TAE 重建指标（PSNR、SSIM、FID）作为感知质量的代理指标并不完善；高运动视频重建是持续性挑战；大规模训练稳定性是主要约束；中文文字渲染是开放领域挑战。

3. 来源材料的接地发现

3.1 视频压缩：基础性权衡

讲座确立了视频压缩是视频生成的关键先决条件，重要性堪比数据质量。两种竞争方法被详细讨论：

Movie Gen 的 TAE（2.5D 方法） 从 2D 图像 VAE 扩展而来，在空间卷积之上添加 1D 时序卷积和时序注意力。可变长度视频通过下采样期间的 StraightenedConv 和上采样期间的最近邻+卷积处理。课程学习从低分辨率逐步推进到高分辨率训练。合成高运动数据通过随机帧采样间隔 1-8 创建，以模拟快速运动。消融研究确定了选择 2.5D 而非 3D 的实际效率原因——3D 在感知指标上略好，但不足以弥补计算成本。

HunyuanVideo 的 3D 因果 VAE 使用 CausalConv3D 共同处理图像和视频，具有因果设计，第一帧单独处理，后续帧自回归处理。它从头训练，使用 L1 + KL + LPIPS + GAN 损失，不使用预训练图像 VAE 初始化。与 Movie Gen 一样，它使用课程学习进行运动密集型视频重建。

这些方法之间未解决的问题——3D 还是 2.5D 时空分解是否真正优越——被明确承认。Movie Gen 出于效率选择 2.5D；3D 在大规模下是否会胜出仍未知。**这个问题仍未解决的原因是，没有已发表的研究在相同训练计算量和数据条件下进行 2.5D 和 3D 设计的控制性消融实验。** 现有证据包括 Movie Gen 的效率导向决策（以质量换取速度）和 HunyuanVideo 的质量验证

方法（从头训练但没有直接比较基线）。这主要是一个评估指标问题：VAE 重建的 PSNR、SSIM 和 FID 指标与下游视频生成质量的感知对齐并不完美，使得难以确定 3D 的边缘质量提升是否会在端到端生成训练后转化为有意义的提升。正确的解决方式需要在相同生成流程中端到端训练两种 VAE 设计，并使用与人类感知对齐的下游指标进行测量——文献中尚未有此类实验报道。

3.2 骨干架构：LLaMA3 与 DiT

Movie Gen 的骨干基于 LLaMA3（仅解码器 Transformer），具有三个关键改进：通过三个串联编码器进行文本条件的交叉注意力层、通过学习参数 γ 和 β 进行时间步/条件注入的 AdaLN 块（替代标准 LayerNorm）、以及由 Flow Matching 非自回归特性启用的全双向自注意力（而非因果注意力）。设计故意镜像 LLaMA3 以最小化训练不稳定性——所有来自 LLM 训练的技术和技巧直接适用。

Movie Gen 的消融研究声称基于 LLaMA3 的架构在质量和文本对齐方面全面优于基于 DiT 的架构。然而，讲座承认了开放问题：在非常大的规模下，LLaMA3 架构是否真正比专用视频架构具有更高的上限仍不明确，因为扩展效应可能掩盖架构差异。**扩展定律研究现已理解扩展可能掩盖架构差异的机制：LLaMA3 和 DiT 风格架构都遵循计算量与损失之间可预测的幂律关系，意味着在足够大的计算预算下，架构之间的质量差距会收敛到评估指标的噪声水平。**一个隔离这种效应的控制实验需要训练多个架构变体在不同扩展计算量级（如 1e18、1e19、1e20 FLOPs），并测量质量差距是否系统性地缩小——这将表明架构等效性——还是保持恒定，这表明真正的架构优势。现有的 Movie Gen 消融在单一规模下进行，不足以区分恒定质量差距和收敛差距。

3.3 文本条件：三种编码器的共识

Movie Gen 的三种文本编码器作为参考设计呈现：MetaCLIP（用于全局语义理解和跨模态对齐的视觉-语言对齐 CLIP 文本塔）、Byte5/BBPE（用于局部字符级特征的字节级分词器，对文本渲染任务至关重要）、以及 UL2（用于另一个全局提示级嵌入的统一语言学习器）。这种多编码器方法比单编码器方法提供更丰富的条件。讲座指出，中文文字渲染是该领域的开放挑战。

3.4 训练：Flow Matching 优于 Diffusion

Flow Matching 通过将生成视为从高斯先验到目标分布的变形来泛化扩散，无需扩散的加噪→去噪约束。Movie Gen 使用 OT Flow Matching，公式 $Xt = (1-(1-\sigma\text{min})t)\cdot x0 + t\cdot x1$，速度预测 $ut = x1 - (1-\sigma min)x0$。消融（表 8）显示 Flow Matching 在整体质量指标上持续优于扩散。推理使用一阶 Euler ODE 求解器和高分辨率视频的分块推理。

3.5 评估结果

Movie Gen 与闭源模型的比较：以微弱优势击败 SORA（单位数净胜率），略微击败 Kling（整体质量净胜率 3.87），但在真实感方面 Kling 表现更好（净胜率-10），在某些维度上以较大优势（净胜率 50-60）超越 Runway 和 Luma。SORA 据称为艺术/电影用例优化，具有高美学和真实感。

3.6 世界模拟器与 AGI 路线图

[未验证声明——讲座来源] 世界模拟器（如物理引擎）被认为是 AGI 的重要基础设施，提供用于合成数据的自我博弈环境和真值反馈。文献提供了 NVIDIA（Cosmos）和 Google DeepMind（Genie 2）在这一方向上积极研究投资的证据，但这是否构成 AGI 的重要基础设施仍未解决。这些系统的详细分析——包括其能力和根本限制——在第 4.2 节主题 5（"世界模拟器与 AGI 路线图"）中展开，该节应被视为本主题的主要参考。

4. 基于文献的深度分析

4.1 保留的详细论文分析

论文 1：Movie Gen：媒体基础模型阵容

arXiv:2410.13720 | Andrew Brown 等（Meta）| 2024–2025

问题： 如何构建统一媒体生成系统，在 30B 参数和 73K 视频词元上下文下具备高质量 1080p HD 视频、同步音频、基于指令的编辑和个性化视频生成能力。

方法： Movie Gen 使用 LLaMA3 解码 Transformer 骨干（30B 参数，73K 词元上下文，16fps 下 16 秒视频），具有三个关键架构改进：（1）用于多编码器文本条件的交叉注意力层（MetaCLIP + Byte5/BBPE + UL2 串联）；（2）来自 DiT 的自适应层归一化（AdaLN）块，通过学习 γ 和 β 进行时间步/条件注入；（3）全双向自注意力（非自回归，与 Flow Matching 兼容）。时序自编码器（TAE）——一种通过添加 1D 时序卷积和时序注意力从 2D 图像 VAE 扩展的 2.5D VAE——在 T/H/W 实现 8×压缩（总计 512×），通道数降至 16。训练使用最优传输 Flow Matching，公式 $x_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$，速度预测 $u_t = x_1 - (1-\sigma_{min})x_0$。

证据： Movie Gen 消融研究（表 8）显示 Flow Matching 持续优于扩散，尤其在整体质量（Q）指标上。基于 LLaMA3 的架构在质量和文本对齐方面全面优于基于 DiT 的架构。与闭源模型比较：以微弱优势击败 SORA；以 3.87 的整体质量略微击败 Kling；在某些维度以较大优势（净胜率 50-60）超越 Runway/Luma。选择 2.5D（TAE）而非 3D 是出于效率考虑，尽管存在边缘感知质量权衡。

相关性： 是讲座关于 LLaMA3 骨干、AdaLN 条件、TAE 视频压缩、OT Flow Matching、三编码器文本条件和双向注意力等技术主张的核心参考。为所有关键架构和训练决策提供了实证验证。

局限性： 未消融 LLaMA3 架构在非常大尺度下是否真正优于专用视频架构。VAE 重建的消融指标（PSNR、SSIM、FID）是感知质量的不完善代理指标。训练不稳定性及细节未披露。

论文 2：HunyuanVideo：大型视频生成模型系统框架

arXiv:2412.03603 | Zijian Zhang 等（腾讯/辉览）| 2024–2025

问题： 如何缩小领先闭源视频生成模型（SORA、Kling 等）与开源社区之间的性能差距，通过构建整合数据策划、架构设计、渐进扩展和高效大规模训练基础设施的全面框架。

方法： HunyuanVideo 使用从头训练的因果 3D VAE（无预训练图像 VAE 初始化），结合 L1 + KL + LPIPS + GAN 损失。因果设计单独处理第一帧，然后自回归处理后续帧。压缩比：T÷4、H÷8、W÷8；通道数降至 16。骨干使用"双流到单流"DiT 架构，扩展 3D RoPE（时间、高度、宽度），并使用预训练多模态 LLM 作为文本编码器。训练使用 Flow Matching 和课程学习（低→高分辨率，合成高运动数据通过随机采样间隔 1-8）。

证据： 13B 参数（最大开源视频生成模型）。专业评估确认其优于 Runway Gen-3、Luma 1.6 和三个表现最佳的中国模型。渐进扩展实现了数百 GPU 稳定训练数周/数月。在足够训练数据下，从头训练的 3D VAE 被验证为有效的。

相关性： 是讲座讨论 HunyuanVideo 架构的主要参考——CausalConv3D 联合处理图像/视频的 3D VAE 设计、运动密集型视频重建的课程学习、Flow Matching 训练。与 Movie Gen 的 2.5D TAE 方法形成对比。

局限性： 与 Movie Gen 在相同基准上没有直接比较。因果设计（单独处理第一帧）与 Movie Gen 的 2.5D 方法不同，但最优选择仍未解决。训练基础设施细节部分专有。

论文 3：Scalable Diffusion Models with Transformers (DiT)

arXiv:2212.09748 | William Peebles, Saining Xie | 2022–2023

问题： Transformer 骨干能否在保持或改善可扩展性的同时替代 U-Net 用于潜在扩散模型？扩散 Transformer 是否表现出与语言模型相同的计算最优扩展定律？

方法： DiT 用操作 VAE 编码潜空间块的视觉 Transformer（ViT）替代 U-Net。关键架构创新：用于条件注入的自适应层归一化（AdaLN）和 AdaLN-Zero。AdaLN-Zero 将缩放因子 γ 初始化为接近零，使残差连接在初始化时不变通过——这对于大规模训练稳定性至关重要。调制网络从条件向量 c 生成六个分量：shift*msa*、*scale*msa、gate*msa（用于注意力）和 shift*mlp、scale*mlp*、*gate*mlp（用于 MLP）。

证据： DiT-XL/2 在 ImageNet 256×256 上达到 FID 2.27——当时扩散模型的 SOTA。更高 Gflops（来自更深/更宽的 Transformer 或更多词元）持续产生更低的 FID，确认幂律扩展。DiT 遵循与语言模型类似的可预测计算扩展，使硬件无关扩展成为可能。

相关性： 是 AdaLN 的基础参考——Movie Gen 和 HunyuanVideo 都使用的关键条件机制。确立了 Transformer 可以在扩散模型中替代 U-Net 并改善可扩展性。AdaLN-Zero 初始化技巧直接应用于所有现代视频生成骨干。

局限性： 专注于图像生成；未涉及视频特有考虑（时间维度、可变长度输入）。该架构为图像设计，后适配到视频，可能留有未探索的优化空间。

论文 4：Flow Matching for Generative Modeling

arXiv:2210.02747 | Yaron Lipman 等（Meta AI）| 2022–2023

问题： 如何在空前规模上训练连续归一化流（CNF）而不需要模拟，以及替代概率路径（超越标准扩散）是否能提高效率和质量。

方法： Flow Matching 训练神经网络通过 ODE 将噪声转换为数据的速度场预测。可与通用高斯概率路径兼容，包括标准扩散路径和最优传输（OT）位移插值。Rectified Flow/OT 路径使用直线插值：X$t = (1-(1-\sigma$min)t)·x$0 + t·x$1，速度 u$t = x$1 − (1−σ$min)x$0。OT 路径提供：（1）更快的训练收敛；（2）更好的样本质量；（3）直线插值路径（对比弯曲扩散路径）；（4）防止单步生成的均值崩溃。使用高斯路径时在数学上与扩散等价。

证据： Flow Matching 在 ImageNet 上训练 CNF，在似然和样本质量上优于替代扩散方法。使用现成 ODE 求解器进行快速可靠采样。OT 路径比扩散路径更高效且泛化更好。

相关性： 为 Movie Gen 和 HunyuanVideo 采用 Flow Matching 而非扩散提供了理论基础。OT 公式 X$t = (1-(1-\sigma$min)t)·x$0 + t·x$1 与讲座公式完全一致。DeepMind 教程也确立了扩散和 Flow Matching 是"同一枚硬币的两面"——在许多场景下可互换。

局限性： 理论框架；未涉及视频特有实现细节（压缩、词元化）。FM 和扩散之间的等价性意味着这种选择通常是务实的而非根本性的。

论文 5：Cosmos World Foundation Model Platform for Physical AI

arXiv:2501.03575 | NVIDIA Cosmos 团队 | 2025

问题： 如何构建统一文本到世界、图像到世界和视频到世界生成的世界基础模型，用于物理 AI 应用，实现合成数据生成、策略评估和机器人与自主系统的闭环仿真。

方法： Cosmos-Predict2.5 使用统一三种生成模态的流式架构。利用 Cosmos-Reason1（物理 AI 视觉-语言模型）进行更丰富的文本接地和更精细控制。在 200M 策划视频片段上训练，使用强化学习后训练。Cosmos-Transfer2.5 是一种 Sim2Real 和 Real2Real 世界转换的控制网风格框架，相比 Cosmos-Transfer1 实现了 3.5 倍模型大小缩减，同时提供更高保真度。

证据： Cosmos-Predict2.5 在 2B 和 14B 规模下相比 Cosmos-Predict1 在视频质量和指令对齐方面实现实质性改进。为机器人策略训练和闭环自主系统仿真实现可靠合成数据生成。

相关性： 直接回应讲座的 AGI 路线图讨论。Cosmos 体现了视频生成与物理 AI 的融合：生成模型成为提供自我博弈环境和代理训练真值反馈的世界模拟器。NVIDIA 的方法（流式统一模型）与 Google 的 Genie 2（自回归潜在扩散）形成对比，但两者追求相同的世界模型基础设施目标。

局限性： 物理 AI 重点意味着较少涉及纯视频质量细节。200M 片段是海量数据集需求。RL 后训练增加了流程复杂性。

论文 6：Genie 2: A Large-Scale Foundation World Model

来源：https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model | Jack Parker-Holder 等（Google DeepMind）| 2024

问题： 如何从单个提示图像生成无限多样的动作可控、可玩的 3D 环境，使 AI 代理能够在多样化的模拟世界中训练和评估，无需手动环境设计。

方法： Genie 2 是一种自回归潜在扩散模型：（1）自编码器将视频编码到潜空间；（2）大型 Transformer 自回归处理潜空间帧；（3）类似 LLM 训练的因果掩码；（4）无分类器引导改善动作可控性。与双向模型（Movie Gen、HunyuanVideo）不同，Genie 2 使用自回归生成以实现实时交互控制。WASD+鼠标输入控制生成的 3D 游戏式环境。

证据： 大规模涌现能力包括：物体交互和可供性、复杂角色动画、物理（水、重力、烟雾）、NPC 行为、照明和反射。世界一致性保持 10-20 秒。SIMA 代理成功在生成环境中遵循自然语言指令。证明视频生成模型可以作为体现 AI 训练的世界模拟器。

相关性： [未验证声明——讲座来源] 验证了讲座关于世界模拟器对 AGI 至关重要的主张——从某种意义上说，Google DeepMind 的 Genie 2 通过 WASD+空格在 3D 游戏式环境中实现代理控制。显示了视频生成与世界模拟在 AI 进步方面的新兴融合。

局限性： 自回归方法以交互可控性换取一定视觉质量。世界一致性限制在 10-20 秒。主要聚焦于游戏式 3D 环境而非逼真视频。

论文 7：Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k

arXiv:2503.09642 | Zangwei Zheng 等（HPC-AI Lab）| 2025

问题： 是否可以以大幅降低的训练成本（20 万美元，比 Movie Gen 或 Step-Video-T2V 低 5-10 倍）实现商业级视频生成质量？哪些技术实现了这一效率突破？

方法： Open-Sora 2.0 使用时空扩散 Transformer（ST-DiT），具有 3D 分块、全时空注意力机制、文本条件交叉注意力和时间步条件 AdaLN。四个维度的联合优化：数据策划、模型架构、训练策略和系统优化。课程学习与渐进分辨率增加和动态批大小。系统优化包括混合精度训练、通信重叠和内存高效注意力。

证据： 根据人类评估和 VBench 分数，Open-Sora 2.0 达到与开源 HunyuanVideo 和闭源 Runway Gen-3 Alpha 相当的性能。20 万美元成本对比同类模型 5-10 倍更高成本，表明了显著的民主化。VBench 评估确认了与领先模型的质量对齐。

相关性： 提供了讲座预测"视频生成质量将随扩展和竞争快速提升；成本将在 1-2 年内大幅下降"的证据。Open-Sora 2.0 体现了这一趋势：通过联合优化以极低成本达到商业级质量。还验证了基于 DiT 的架构（ST-DiT）与基于 LLaMA3 的方法仍具有竞争力。

局限性： 20 万美元涵盖训练计算但不涵盖基础设施开发成本。VBench 分数提供比较但不完整披露技术基准。同时需要四个优化维度的专业知识才能实现效率提升。

论文 8：Scaling Laws For Diffusion Transformers

arXiv:2410.08184 | Zhengyang Liang 等（腾讯 AI Lab、上海 AI Lab）| 2024

问题： 扩散 Transformer 是否遵循可预测的计算幂律扩展关系，从而能够对给定训练预算下的最优模型大小和数据需求进行精确预测？

方法： 在 DiT 上跨 1e17 至 6e18 FLOPs 计算预算进行系统实验。确认幂律关系：Loss ~ Compute^(-α)，其中 α 是扩展指数。关键发现：预训练损失与生成质量（FID）相关，使廉价预训练损失成为质量代理。发现视频扩散模型对学习率和批大小比语言模型更敏感——这是一个需要显式建模的关键差异。

证据： DiT 在所有模型大小和分块维度上遵循可预测的计算扩展。预训练损失是 FID 的有效代理，使训练决策无需完整评估。推算能力：给定计算预算预测最优模型大小和数据需求，或预测 1B 参数和 1e21 FLOPs 下的训练损失。

相关性： 验证了讲座关于"扩展效应占主导"的观点。扩展定律的存在使训练更具工程可预测性。然而，视频特定超参数敏感性意味着架构选择与扩展以非平凡方式交互。DiT 与 LLaMA3 架构辩论可能通过扩展解决——如果两者都遵循相似幂律。

局限性： 专注于图像 DiT；由于时间维度和更长训练运行，视频特定扩展可能不同。计算最优预测假设固定架构——不同规模下的最优架构尚未确定。

论文 9：Towards Precise Scaling Laws for Video Diffusion Transformers

arXiv:2411.17470 | Yuanyang Yin 等 | 2024

问题： 视频扩散 Transformer 如何随计算扩展，以及除对语言模型关键的超参数外还有哪些超参数必须建模？

方法： 对视频扩散 Transformer 跨不同模型大小和计算预算进行扩展定律系统分析。发现视频扩散模型对学习率和批大小比语言模型更敏感。提出预测最优超参数（LR、BS）的新扩展定律，适用于任何模型大小和计算预算。在 1e10 TFlops 预算范围内验证，相比常规扩展方法实现 40.1%推理成本降低。

证据： 视频扩散 Transformer 存在幂律关系。最优 LR = f(模型大小, 计算)和最优 BS = g(模型大小, 计算)需要显式建模。广义损失-计算关系实现非最优模型大小的性能预测——对推理受限场景有用。

相关性： 通过添加视频特定敏感性分析完善了论文 8 的扩展定律讨论。解释了"大规模训练稳定性是主要约束"的原因——视频扩散模型需要语言模型所不需要的仔细逐规模超参数调优。通过最优超参数实现 40.1%推理成本降低具有重要实际意义。

局限性： 在特定计算预算范围（1e10 TFlops）内验证；向更大规模（Movie Gen 的 30B 模型）推算不确定。需要广泛消融为每个新模型规模确定最优超参数。

论文 10：DC-VideoGen: Efficient Video Generation with Deep Compression Video Autoencoder

arXiv:2509.25182 | Han Cai 等（NVIDIA）| 2025

问题： 如何通过激进的潜空间压缩为预训练视频扩散模型实现显著的推理效率提升，同时不牺牲质量？

方法： DC-VideoGen 使用深度压缩视频自编码器，采用新颖的分块因果时序设计，实现 32×/64×空间和 4×时间压缩（对比 Movie Gen 的 8×/8×/8×）。AE-Adapt-V 适应策略使预训练模型能够快速稳定地转移到新的深度压缩潜空间，Wan-2.1-14B 仅需 10 个 GPU 天（H100）。

证据： 推理延迟比基线模型降低高达 14.8 倍，且无质量妥协。实现单 GPU 生成 2160×3840 视频（此前需要多 GPU）。后训练加速方法适用于任何预训练视频扩散模型。

相关性： 为讲座的效率和成本预测提供了直接证据。14.8 倍延迟降低和单 GPU 高分辨率生成展示了视频生成的快速民主化。深度压缩（32×/64×空间）扩展了讲座讨论的 TAE（8×/8×/8×）和 3D VAE（8×/8×/4×）的压缩比与重建质量之间的权衡。

局限性： 后训练适应需要访问原始预训练模型权重。分块因果设计可能并非所有生成场景最优。最大压缩（64×空间）的质量在规模上需要进一步验证。

论文 11：VBench: Comprehensive Benchmark Suite for Video Generative Models

arXiv:2311.17982 | Yuming Jiang 等（上海 AI Lab、NVIDIA 等）| 2023

问题： 如何用与人类感知对齐的可分解质量维度系统评估视频生成模型，为未来发展提供可操作的见解？

方法： VBench 将视频生成质量分解为 16 个分层维度：主体质量（身份一致性、美学质量、成像质量）、运动与时序（平滑度、闪烁、幅度）、空间与视觉关系、文本对齐和风格一致性。每个维度使用定制提示和评估方法以及人类偏好注释进行对齐验证。

证据： 跨所有 16 个维度对多个模型的综合评估揭示了各模型的优势和劣势。人类对齐验证确保自动指标反映感知质量。VBench 已成为现代视频生成研究的标准评估框架。

相关性： 提供了 Open-Sora 2.0、VideoAR 和大多数现代论文用于质量比较的系统评估框架。讲座关于"大幅超越 Runway"和"Flow Matching 优于扩散"的主张基于此类基准评估。

局限性： 16 个维度仍无法涵盖视频质量的所有方面。人类偏好注释成本高昂，可能无法随快速模型发展扩展。某些维度可能具有固有评估噪声。

论文 12：Temporal Regularization Makes Your Video Generator Stronger

arXiv:2503.15417 | Haodong Chen 等 | 2025

问题： 如何通过数据级干预（无需架构修改）提高视频生成的时序一致性和多样性？

方法： FluxFlow 在数据级应用受控时序扰动：时序窗口内帧洗牌、时序 dropout（随机帧掩码）和时序速度变化。无需架构更改——作为数据预处理增强使用。正则化效果减少对特定时序模式的过拟合，提高对未见运动的泛化。

证据： FluxFlow 在 UCF-101 和 VBench 基准上显著改善 U-Net、DiT 和 AR 架构的时序一致性和多样性，同时保持空间保真度。跨架构的普遍改进表明时序正则化是基本属性而非架构特有。

相关性： 回应了讲座关于"高运动视频重建仍具挑战"的约束。FluxFlow 提供与 Movie Gen 课程学习互补的原则性数据级增强（随机采样间隔 1-8 用于合成高运动数据）。两种方法都认识到训练数据中时序多样性对质量至关重要。

局限性： 时序增强增加训练复杂性，可能需要仔细调整扰动强度。最优增强策略可能因架构而异。

论文 13：VideoAR: Autoregressive Video Generation via Next-Frame & Scale Prediction

arXiv:2601.05966 | Junyuan Shang 等 | 2026

问题： 自回归（AR）模型能否在提供卓越推理效率的同时与扩散/流模型在视频生成质量上竞争？

方法： VideoAR 是首个大规模视觉自回归（VAR）视频生成框架。结合多尺度下一帧预测与自回归建模，通过帧内 VAR 建模和因果下一帧预测解耦空间和时序依赖。使用 3D 多尺度分词器实现高效时空编码。长期一致性的三个创新：多尺度时序 RoPE（将旋转位置嵌入扩展到时间维度）、跨帧误差纠正（通过未来上下文检测和纠正累积误差）和随机帧掩码（缺失帧的训练正则化）。

证据： UCF-101 上 FVD：99.5 → 88.6（AR 模型中的 SOTA）。VBench 分数：81.74——与比其大 10 倍的扩散模型相当。推理步骤比扩散减少 10 倍以上。多阶段预训练与课程方法（低→高分辨率和时长）逐步对齐空间和时序学习。

相关性： 挑战了讲座关于扩散/流模型是确定路径的隐含假设。表明 AR 模型可以以卓越效率（减少 10 倍推理步骤）达到相当质量。架构辩论（LLaMA3 vs DiT vs AR）仍未解决——VideoAR 证明了 AR 的可行性。跨帧误差纠正解决了 AR 的误差累积弱点。

局限性： 尽管有缓解策略，自回归误差传播仍是根本挑战。长期一致性（10+秒）可能仍落后于双向流模型。

论文 14：OmniWeaving: Towards Unified Video Generation with Free-form Composition and Reasoning

arXiv:2603.24458 | Kaihang Pan 等 | 2026

问题： 如何弥合专有全能视频生成系统（如 Seedance-2.0）与碎片化开源替代方案之间的差距，在单一开源框架中实现统一文本+图像+视频+推理生成？

方法： OmniWeaving 将文本到视频、多图像到视频、视频到视频编辑和推理增强生成集成在一个框架中。使用大规模预训练与多样化组合和推理增强场景。学习在推断复杂用户意图的同时临时绑定交错的文本、多图像和视频输入。引入 IntelligentVBench 基准用于下一代智能统一视频生成评估。

证据： 开源统一模型中的 SOTA 性能。与专有 Seedance-2.0 相当。IntelligentVBench 基准为推理增强视频生成建立评估框架。

相关性： 代表视频生成研究的前沿——向统一模型的架构收敛。为 AGI 路线图趋势提供证据：推理增强训练将视频生成与"代理 AI"阶段联系起来。确认讲座讨论的竞争格局：开源模型正在追赶，但在统一/全能系统方面仍有差距。

局限性： 统一模型复杂性增加训练和推理成本。推理增强训练需要专业数据集和评估。相比任务特定模型仍处于早期阶段。

论文 15：VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models

arXiv:2411.13503 | Yuming Jiang 等 | 2024

问题： 如何将 VBench 扩展为更全面的评估框架，覆盖多种视频生成任务（文本到视频、图像到视频）以及信任度评估以进行整体性能评估？

方法： VBench++跨 16 个维度扩展 VBench，使用自适应图像套件在不同设置下进行公平比较。添加结合多个质量维度的信任度评估。支持文本到视频和图像到视频评估，各有专用提示和指标。

证据： 扩展框架实现不同视频生成场景和模型类型之间的公平比较。信任度评估提供超越单个维度分数的整体质量评估。

相关性： 为大多数现代视频生成论文提供系统评估方法。讲座的基准比较依赖 VBench++等框架进行定量验证。

局限性： 扩展框架增加评估复杂性和成本。个体维度之间的权衡可能无法被聚合信任度分数完全捕获。

4.2 综合主题评估

主题 1：架构收敛与未解决的骨干辩论

文献揭示了视频生成向解码 Transformer 骨干的明确趋势，但三种不同架构范式已出现，各有权衡：

基于 LLaMA3 的方法（Movie Gen）：30B 参数 LLaMA3 骨干继承所有 LLM 训练稳定性技术。其三个改进（用于多编码器条件的交叉注意力、用于时间步注入的 AdaLN、用于非自回归 Flow Matching 的双向自注意力）代表了对成熟 LLM 基础设施的务实复用。Movie Gen 的消融确认了在质量和文本对齐方面全面优于基于 DiT 的方法，但该消融可能未充分考虑规模效应。

基于 DiT 的方法（HunyuanVideo、Open-Sora 2.0）："双流到单流"带 3D RoPE 的 DiT 架构和带时空注意力的 ST-DiT 代表同一范式的两个变体。Open-Sora 2.0 的 20 万美元结果表明基于 DiT 的架构可以以大幅降低的成本达到与基于 LLaMA3 方法相当的质量。DiT 扩展定律论文确认了可预测的幂律关系，表明架构差异在大规模下变得不那么重要。

自回归方法（VideoAR、Genie 2）：VideoAR 的 81.74 VBench 分数——与比其大 10 倍的扩散模型相当——以及 10 倍推理步骤减少代表了一种根本不同的效率画像。Genie 2 的自回归潜在扩散模型优先考虑交互可控性而非原始质量。这些结果表明 AR 与扩散/流的辩论并未解决，可能很大程度上取决于目标用例（批量生成 vs. 交互控制）。

目前最强证据支持基于 LLaMA3 的方法用于最高质量（根据 Movie Gen 消融），基于 DiT 的方法用于效率和民主化（根据 Open-Sora 2.0），以及 AR 方法用于推理效率和交互应用（根据 VideoAR）。DiT 扩展定律表明三种范式在大规模下可能收敛。

主题 2：视频压缩——2.5D 与 3D 权衡及深度压缩的出现

文献讨论的三种视频压缩方法（TAE 2.5D、3D 因果 VAE 和深度压缩 VAE）代表了效率-质量权衡的光谱：

2.5D TAE 方法（Movie Gen）实现 8×/8×/8×压缩，通道降至 16，出于效率选择 3D。3D 因果 VAE 方法（HunyuanVideo）实现 T÷4、H÷8、W÷8 相同通道压缩，从头用多分量损失训练。关键洞察是两种方法都同意联合图像-视频训练至关重要，且需要某种形式的时序建模超越逐帧压缩。

REDUCIO（arXiv:2411.13552，2024） 通过在图像条件 VAE 中利用帧间冗余实现 64×潜空间压缩，进一步扩展了深度压缩轨迹。这使得单块 A100 GPU 上 16 秒内生成 1K 视频成为可能——一个显著的效率里程碑。REDUCIO 的方法与 DC-VideoGen 的分块因果设计互补：两者都追求极端压缩但通过不同机制（图像条件潜空间利用 vs. 分块级因果建模）。64×压缩

超过 DC-VideoGen 的空间压缩范围（32×/64×），表明适当设计下当前模型使用的压缩比以外的空间可能仍可行。

ARVAE（arXiv:2512.11293，2025） 通过下采样光流进行运动建模和空间补偿新内容的解耦时空表示，提出作为 TAE 和 CausalConv3D 方法的互补替代方案。这种解耦设计允许时序一致性和空间保真度的独立优化——这是一种结构选择，而 2.5D TAE（通过共享卷积耦合时空）和 CausalConv3D（通过因果自回归耦合）都未显式支持。ARVAE 证明通过这种架构替代可以实现优越的重建质量，表明该领域尚未收敛到视频压缩中时空信息分离和重组的最优方式。

REDUCIO 和 ARVAE 共同说明视频压缩设计空间仍在积极探索中，互补方法针对不同方面：REDUCIO 推向极端压缩比，而 ARVAE 完善表示分解。DC-VideoGen 在更深压缩方面的成功表明该领域正在向更好网络的更深压缩发展，使 2.5D 与 3D 的区别可能不如压缩设计本身的质量重要。

主题 3：Flow Matching 作为主导训练目标

Flow Matching 已在所有主要视频生成系统中获得广泛共识。文献同时提供了理论论证（论文 4：OT 路径提供更快收敛、更好质量和直线插值）和实证验证（Movie Gen 表 8 消融确认 FM 持续优于扩散）。使用高斯路径时与扩散的数学等价性意味着实践者可以将此选择视为务实而非根本性的。

最优传输公式值得特别关注：直线插值路径 $x_t = (1-(1-\sigma_{min})t)\cdot x_0 + t\cdot x_1$ 与速度预测 $u_t = x_1 - (1-\sigma_{min})x_0$ 产生比弯曲扩散路径更高效的训练和更好泛化。单步生成防止均值崩溃对推理效率特别重要。这使 OT Flow Matching 成为文献对新视频生成项目最明确的建议。

然而，VideoAR 在自回归建模上的成功挑战了 Flow Matching 普遍最优的隐含假设。AR 方法以根本不同的生成机制达到相当的 VBench 分数，表明训练目标和生成范式是可分离的设计决策。

主题 4：效率与民主化轨迹

文献记录的成本轨迹引人注目。Movie Gen 和 Step-Video-T2V 需要估计数亿美元的计算预算。Open-Sora 2.0 以 20 万美元达到相当质量——代表 5-10 倍降低。DC-VideoGen 在此基础上增加 14.8 倍推理加速。这些结果共同表明，在 1-2 年内，商业级视频生成可能对个人研究人员和小团队触手可及。

实现这一民主化的机制多种多样：数据策划、模型架构、训练策略和系统优化的联合优化（Open-Sora 2.0）；将推理延迟降低一个数量级的深度压缩 VAE（DC-VideoGen）；以及一般扩展定律洞察，即用适当超参数在更多数据上训练的更小模型可以匹配更大模型（论文 8 和 9）。通过最优超参数预测实现 40.1%推理成本降低（论文 9）为效率提升增加了另一个维度。

这一轨迹验证了讲座的预测："视频生成质量将随扩展和竞争快速提升；成本将在 1-2 年内大幅下降。"实际含义是，视频生成研究的进入壁垒正在快速降低，大实验室的竞争护城河正从原始计算转向数据质量、评估方法论和特定应用优化。

主题 5：世界模拟器与 AGI 路线图

视频生成与世界模拟的融合现在是 NVIDIA（Cosmos）和 Google DeepMind（Genie 2）的明确研究方向，OmniWeaving 推向推理增强统一模型。文献支持这一方向的积极投资：

- **Cosmos（NVIDIA）**：流式世界基础模型，在 2B 和 14B 规模统一 Text2World、Image2World、Video2World，在 200M 策划片段上训练并使用 RL 后训练。为机器人策略训练和自主系统闭环仿真实现合成数据生成、策略评估和闭环仿真。
- **Genie 2（Google DeepMind）**：自回归潜在扩散模型，生成动作可控 3D 环境。在 10-20 秒内保持世界一致性，SIMA 代理成功在生成环境中遵循自然语言指令。
- **OmniWeaving**：作为前沿的推理增强统一视频生成，推理增强训练将视频生成与代理 AI 阶段联系起来。引入 IntelligentVBench 作为推理增强生成的评估框架。

然而，文献也揭示了根本限制。Genie 2 和 Cosmos 学习的视觉相关性逼近物理，而非显式物理定律。在没有物理验证的情况下将这些作为体现 AI 的训练环境使用，可能导致策略利用视觉伪影而非学习真正的物理理解。在大多数情况下，世界一致性在 10-20 秒后下降，物理准确性（精确轨迹、碰撞响应）无法保证。**合成数据在世界模拟器中引入系统性偏差的机制通过三种复合途径运作：（1）分布偏移——生成环境在分布上必然与真实训练数据不同，导致在模拟器中训练的策略泛化到真实环境时不完美；（2）复合误差传播——当视频生成模型用于闭环训练循环时，其输出训练影响下一批生成的策略，物理近似的误差累积而非平均，导致系统性偏差的物理；（3）伪影利用——由于生成的物理学习视觉相关性而非强制执行守恒定律，训练后的策略可以发现并利用视觉伪影（例如视觉逼真但物理上不可能的物体交互），这些在真实环境中不存在。** 这意味着来自视频生成模型的世界模拟器最适合预训练或多样性增强，而非作为体现代理的唯一训练环境。世界模拟器的融合是真实的且正在加速，但通过视频生成通往可靠 AGI 基础设施的道路受到这些根本限制的约束。

主题 6：评估成熟度与质量评估标准化

VBench 和 VBench++已建立 16 个与人类感知对齐的可分解维度的标准化评估框架。这一评估成熟度对模型质量的基于证据的主张至关重要——讲座关于"大幅超越 Runway"和"Flow Matching 优于扩散"的主张依赖这些框架进行定量验证。

评估文献中最重要的洞察是整体指标（FVD、FID、CLIPSIM）与人类判断不一致。这验证了可分解评估的必要性：不同模型在不同维度上表现优异，聚合指标可能掩盖特定弱点。VBench++通过添加信任度维度（文化公平、性别偏见、肤色偏见、安全性）扩展了这一点，随着视频生成部署到消费应用变得越来越重要。

OmniWeaving 的新兴 IntelligentVBench 将评估扩展到推理能力，反映了该领域认识到下一代视频生成需要评估组合推理、时空因果性和多模态整合——当前基准未涵盖的维度。

5. 当前项目的综合评估

基于对来源材料和 15 篇已开放论文的综合，以下综合评估浮现：

视频生成领域已达到若干设计决策有证据支持的成熟度。OT Flow Matching 是推荐的训练目标——理论优势（更简单公式、直线路径、更快收敛、无模式崩溃）得到 Movie Gen 实证消融研究验证，且与扩散的数学等价性意味着不牺牲表达能力。根据扩展定律研究，LLaMA3 或 DiT 基础的 Transformer 骨干是数据质量和计算之后的次要选择。多编码器文本条件（MetaCLIP + BBPE + UL2 或等效）应作为标准。AdaLN 条件与 AdaLN-Zero 初始化在所有现代方法中通用，对大规模训练稳定性至关重要。FluxFlow 风格时序增强是一种低成本架构无关的添加，可改善跨架构的时序质量。

来源讲座中的若干假设需要修改或完善。关于扩散/流模型明确优于自回归方法的隐含假设受到 VideoAR 竞争性 VBench 分数和 10 倍推理效率优势的挑战。2.5D 与 3D VAE 的辩论是真实的且经验上未解决——但 REDUCIO 和 ARVAE 表明压缩设计空间正在多个方向积极扩展，表明 2.5D 与 3D 之间的选择是更广泛设计优化问题的的一个方面而非二元决策。通过世界模拟器的 AGI 路线图得到方向性支持，但鉴于视觉相关性与显式物理的根本限制以及上述合成数据偏差机制，应被视为长期研究方向，而非近期能力声明。

新项目最不确定的领域是：目标规模下的最优骨干架构（LLaMA3 vs DiT vs AR）；目标用例的最优压缩比和设计；以及目标模型规模的特定超参数敏感性景观。这些不确定性可以通过针对性消融研究解决，使它们成为研究机会而非障碍。

6. 未解决问题与关键决策缺口

1. 目标规模下 LLaMA3 vs DiT vs AR 骨干优越性。 Movie Gen 消融声称 LLaMA3 > DiT，但 VideoAR 证明 AR 可以用减少 10 倍推理步骤匹配扩散。架构层面的辩论未解决——扩展效应可能主导任何一方。需要多架构控制在不同计算规模（1e18 至 1e20 FLOPs）的消融，测量质量差距是否收敛（表明架构等效）还是保持恒定（确认真正架构优势）。任何具体项目都需要在相同计算和数据下进行比较研究来澄清。

2. 目标规模下的最优视频压缩设计。 压缩比之间的权衡（更多压缩=更少词元=更快但可能更低质量）是领域相关的。REDUCIO（64×潜空间压缩）和 DC-VideoGen 的（32×/64×空间压缩）表明比当前标准更深得多的压缩是可行的，而 ARVAE 表明时空信息分解方式（通过光流解耦）可能与原始压缩比同样重要。2.5D 与 3D 权衡是这一更广泛问题的一个具体实例。

3. 非常大尺度视频模型的超参数扩展。 视频 DiT 扩展定律（论文 9）显示 LR/BS 敏感性不同于语言模型，需要逐规模调整。然而，30B 规模（Movie Gen 的规模）模型的最优超参数计划尚未充分描述。这直接影响训练稳定性——这是讲座的关键约束之一。

4. 长期世界一致性与物理准确性。 Genie 2 在 10-20 秒（最长 1 分钟）内保持一致性。用于代理训练的扩展仿真（数分钟到数小时）的更长视频生成尚未展示。误差累积和长期世界一致性退化是约束世界模拟器用例的根本挑战。合成数据偏差机制（分布偏移、复合误差、伪影利用）使体现 AI 应用的这一限制更加复杂。

5. 视频生成中的中文文字渲染。 在讲座和文献中都被明确标识为开放挑战。15 篇已开放论文中没有一篇提供了令人满意的解决方案。这是一个影响中文市场部署的特定能力缺口。

6. 压缩、骨干和文本条件之间的最优计算分配。 文献为骨干组件提供了扩展定律，但没有充分理解在所有三个组件之间固定计算预算的最优分配。这是任何新项目的实际关键决策问题。

7. 建议的后续步骤

1. 在目标规模下进行针对性消融，比较基于 LLaMA3 与基于 DiT 的骨干。 Movie Gen 现有消融提供了参考点，但在其特定规模和特定数据下进行。需要在目标应用和计算预算下确定最优骨干选择的项目特定消融。消融应同时测量质量（VBench）和效率（推理延迟、内存占用）。

2. 针对目标用例评估深度压缩 VAE 方法。 DC-VideoGen 的 32×/64×空间压缩分块因果设计和 REDUCIO 的 64×图像条件方法代表了与讲座讨论的 2.5D/3D 方法根本不同的效率-质量权衡。在目标预训练模型上运行 AE-Adapt-V 适应实验将澄清深度压缩对特定应用是否有益。

3. 将 FluxFlow 风格时序增强纳入训练流程。 这是一种低成本、架构无关的数据级添加，可改善 U-Net、DiT 和 AR 骨干的时序质量。特定扰动参数（帧洗牌窗口大小、dropout 率、速度变化范围）应根据特定数据领域进行调整。

4. 使用 VBench++作为主要框架设计综合评估计划。 鉴于整体指标（FVD、FID）与人类判断不一致的证据，评估计划应使用 VBench++维度作为主要质量指标，并辅以与目标应用相关的领域特定维度。这避免了在聚合指标掩盖特定弱点下优化的陷阱。

5. 评估 OmniWeaving 的 IntelligentVBench 基准框架作为目标应用的潜在评估协议。 IntelligentVBench 将标准视频生成评估扩展到推理增强生成，覆盖组合推理、时空因果性和多模态整合。即使对此框架进行初步评估——无需完整实施——也将澄清目标应用是否需要推理能力，并识别统一多模态视频生成系统的竞争格局。

6. 专门针对视频生成中的中文文字渲染进行文献综述，按三个搜索维度组织： （a）字体特定生成方法——用于在视频输出中生成特定字符字形和排版样式的专用神经渲染方法；（b）OCR 引导视频文字渲染——首先通过 OCR 或布局预测生成文字区域，然后以受控字体、大小和样式渲染字符序列的管道方法；（c）从文本到图像的跨语言文字渲染迁移——利用处理多语言文字的文本到图像渲染进展（例如 AnyText、TextDiffuser），并研究这些技术是否迁移到视频领域。跨这三个维度的结构化搜索将澄清哪种方法对特定应用上下文最有前景。

8. 关键风险、注意事项和证据边界

1. 世界模型物理准确性从根本上局限于视觉相关性。 Genie 2 和 Cosmos 学习逼近物理的视觉相关性，而非显式物理定律。在生成的世界模拟器中训练的策略可能利用视觉伪影而非学习真正的物理理解。任何将视频生成作为世界模拟器基础设施的应用都应包括显式物理验证，而不仅仅是视觉质量指标。

2. 与闭源模型的比较主张基于有限信息。 SORA、Kling 和其他闭源模型未披露训练细节、基准方法论或评估协议。所有比较主张（包括 Movie Gen"微弱超越 SORA"）都基于有限的公开披露，应被视为近似方向信号而非精确测量。

3. 超参数敏感性意味着扩展并非简单。 视频扩散模型对学习率和批大小的敏感性显著高于语言模型。论文 8 和 9 的扩展定律提供了有用的框架，但需要逐规模超参数调整。跳过此调整步骤的项目面临训练不稳定性或次优质量的风险。

4. 评估框架成熟度并未扩展到所有感兴趣维度。 VBench 和 VBench++对标准文本到视频质量维度进行了良好验证，但世界模型评估（物理准确性、因果一致性、长期动力学）、推理评估（组合理解、反事实生成）和领域特定质量维度尚未标准化。面向这些维度的项目必须设计自定义评估协议。

5. 训练计算估计不包括基础设施开发成本。 Open-Sora 2.0 的 20 万美元数字涵盖训练计算，但不涵盖构建训练系统所需的基础设施开发、工具和工程努力。实际项目成本可能显著高于仅计算成本估计。

6. 从图像到视频领域的证据迁移风险。 多篇基础论文（DiT、Flow Matching）专注于图像生成，未涉及视频特有考虑。这些机制在实践中有良好迁移（AdaLN、OT Flow Matching），但视频特有适应（时序一致性、可变长度处理、运动动力学）可能揭示图像领域不存在的失效模式。

7. 合成数据偏差传播风险。 FluxFlow 的时序增强和世界模型在合成视频上的训练都可能通过三种复合机制引入系统性偏差：生成环境和真实环境之间的分布偏移；训练循环中误差累积而非平均的复合误差传播；以及策略学习视觉捷径而非真正物理理解的伪影利用。使用合成数据的项目应包括多样性和准确性验证协议，不应依赖纯合成训练用于体现代理应用。