**视频生成模型的预训练和模型架构分析：**

**以Sora、MovieGen、HunyuanVideo为例**

**1120232535 汪隽宁 数据科学与大数据技术**

# 引言

随着人工智能生成内容（AIGC）技术的迅速发展，视频生成模型已成为多模态领域的研究热点之一。相比于图像或文本生成，视频生成任务在时空建模、连贯性保持、物理规律遵守等方面提出了更高的挑战。近年来，随着大规模预训练模型在语言、图像领域的成功，研究者开始尝试将类似的技术范式迁移至视频生成任务，推动了这一领域的快速进步。

在众多视频生成系统中，OpenAI 提出的 Sora 模型，以其惊人的视觉质量与长时长视频建模能力，引发了学术界与工业界的广泛关注。Sora 的技术细节虽未完全公开，但从其技术报告与演示结果中可窥见其在视频压缩表示、空间-时间建模及高效采样等方面的创新[1]。与此相比，Meta提出的 MovieGen 模型和腾讯混元提出的 HunyuanVideo 模型，作为代表性的视频生成方案，有和Sora大体相近的整体思路，但在很多方面的细节处理上有自己的特色[2][3]。

本调研报告以 Sora、MovieGen 和 HunyuanVideo 三个模型为案例，从预训练、模型架构、模型评估指标以及未来发展趋势等方面展开分析，旨在梳理当前视频生成模型的主流技术路径与关键挑战，并对视频生成的未来发展进行展望。

# 预训练

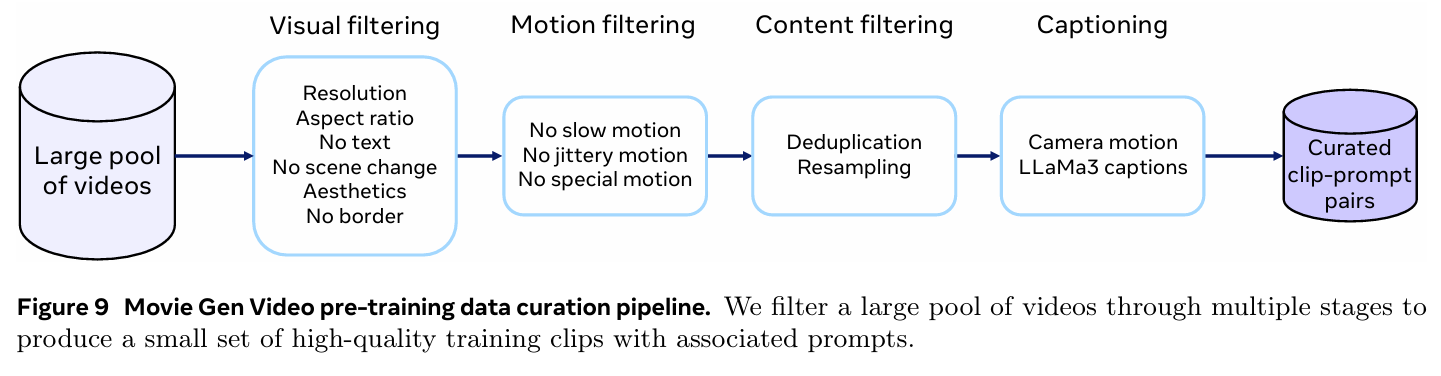
在生成式AI甚至整个机器学习领域中，数据都是至关紧要的一环。很多情况下，数据质量的好坏往往决定了最后训练出的模型的性能。

本部分将主要聚焦于视频生成模型的预训练数据收集、预训练数据处理、预训练阶段本身等方面进行分析。Sora 技术报告在这一部分的内容十分简略，有效信息较少，故本部分主要基于MovieGen和HunyuanVideo的论文内容进行分析。

## 预训练数据收集与处理

在Sora的技术报告中，仅说明其使用了三种数据集进行训练，包括：公开数据，主要收集自行业标准的机器学习数据集和网络爬虫；通过合作伙伴获取的专有数据，例如OpenAI与 Shutterstock⁠ Pond5 合作，构建并交付 AI 生成的图像；内部开发的自定义数据集，包括来自 AI 培训师、团队成员和员工的反馈。

MovieGen论文中的内容则极为详实。其预训练数据集包含O(100)M个视频-文本对和O(1)B个图像-文本对。其原始数据池包含时长 4 秒到 2 分钟的视频，涵盖人类、自然、动物和物体等不同领域的概念。在经过数据处理流程后最终的预训练片段-提示对集中，每个片段时长 4 秒到 16 秒，采用单镜头拍摄，且不包含细微动作。这样的训练视频对于模型预训练更为简单，比较友好，其他的视频生成模型大多也对预训练数据进行了相似的处理。



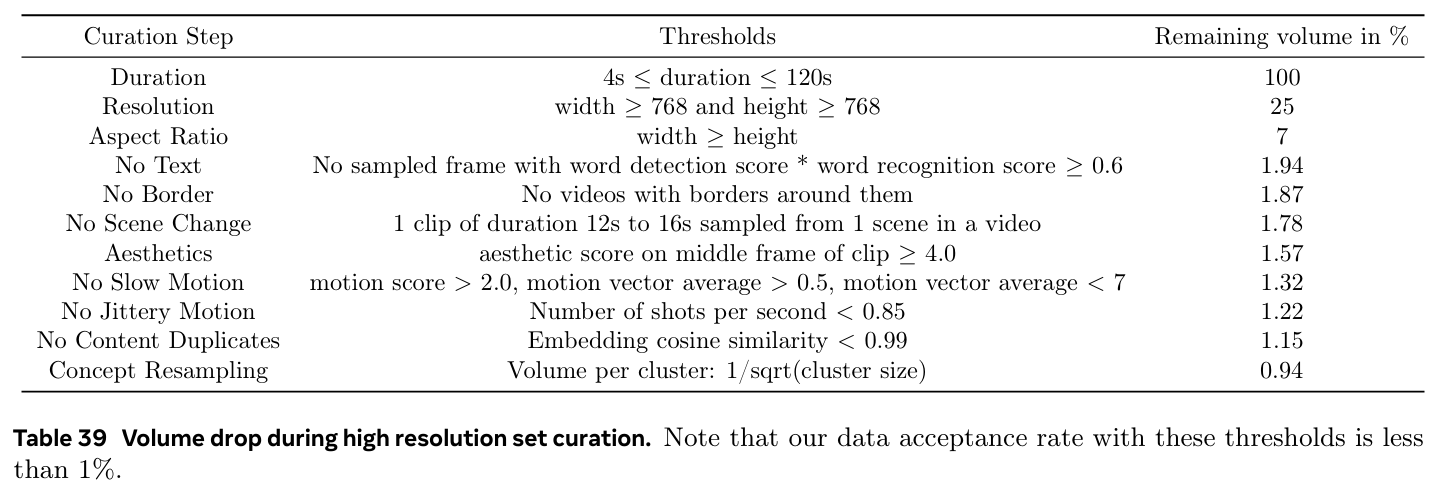
在图文数据预处理方面，MovieGen遵循与Meta先前工作[4]类似的策略，我们主要关注视频数据预处理。其视频数据处理流程如上图所示。它包含三个过滤阶段：视觉过滤，运动过滤，内容过滤，以及文字重新匹配阶段。过滤后的片段会生成详细的注释文字，平均包含 100 个单词。我们接下来分析每个阶段的内容和作用。

**视觉过滤** MovieGen使用了一系列性能较好的现成的基于模型的过滤器筛选掉低视觉质量的视频，先移除长宽像素过少的视频，保留的视频中横屏占60%，竖屏占40%。接下来删除了文字过多的视频，分割了分镜过多的视频。最后进行了美学层面的过滤，并对有边框的视频进行了删除。

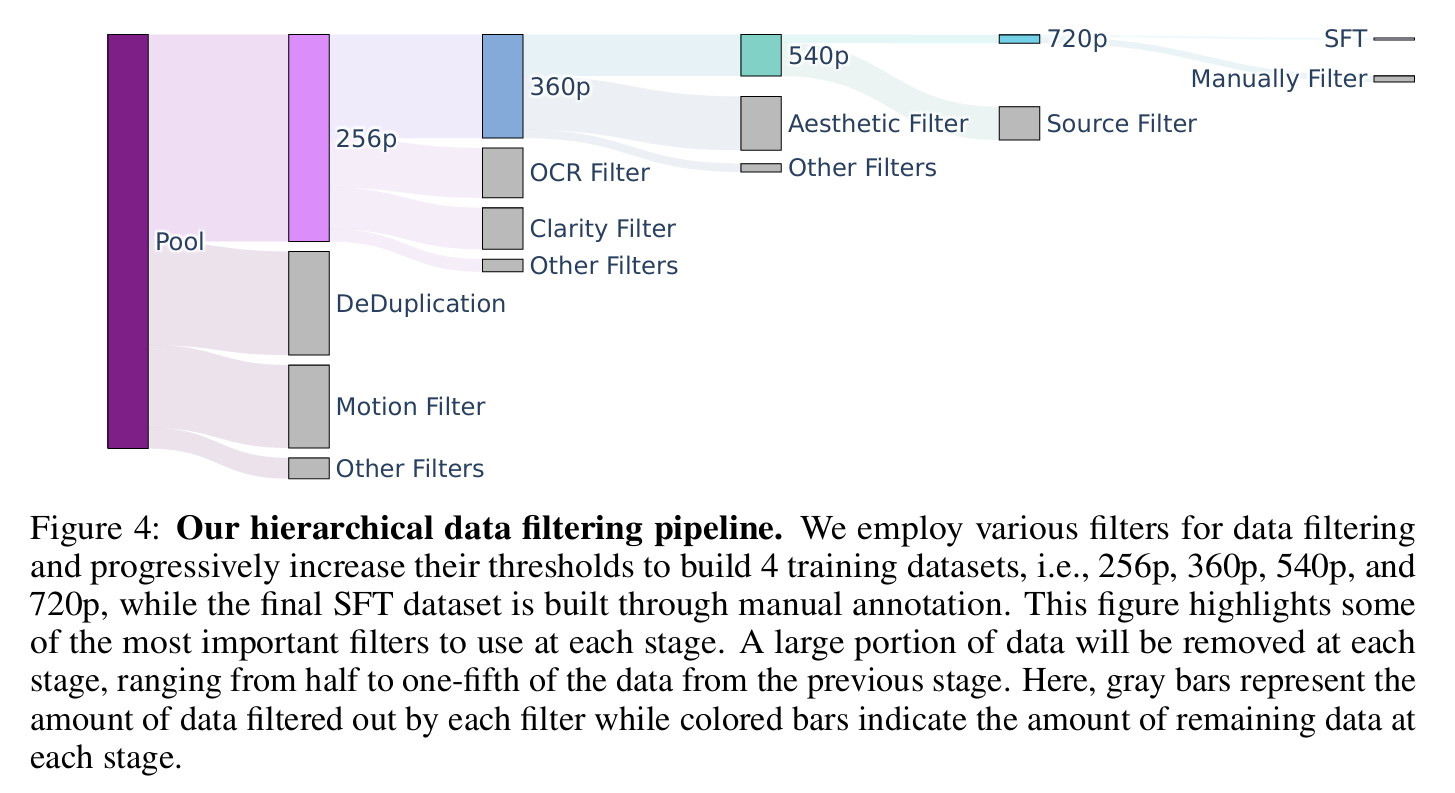
**运动过滤** MovieGen遵循先前的研究[5]自动滤除低运动视频，并使用内部静态视频检测模型移除无运动的视频。接下来，他们识别出具有“合理”运动的视频。最后移除了带有特殊运动效果的视频，例如幻灯片视频。

**内容过滤** 为了确保预训练集的多样性，MovieGen使用复制检测嵌入空间[6]中的相似性进行去重。此外，他们还使用了聚类方法，将重复的聚类合并之后重新采样。

**文字重新匹配** MovieGen微调了LLaMa3-Video模型的 8B 和 70B 版本，为视频片段创建准确且详细的文本提示。此外，为了实现电影级的摄像机运动控制，他们训练了一个摄像机运动分类器，它可以预测 16 种摄像机运动类别。



经过以上一系列视频数据预处理，如上表所示，MovieGen最后剩余的视频量不到原来的1%，可见数据处理这一部分的工作量是极大的。



如上图所示，HunyuanVideo在预训练阶段的数据处理流程和MovieGen相似，我们这里就对一些新内容进行分析。在预处理阶段，混元使用了OpenCV中的拉普拉斯算子来找到视频中最清晰的一帧，将其作为开始帧。此外，在数据清洗这一阶段，很难保证某一步的数据清洗对于训练是有益还是有害，混元通过先在一个小规模的模型上进行实现，验证其清洗步骤对于模型性能的增益，这是一个很有新意的方法。

## 预训练

这一部分我们从MovieGen的预训练开始看起。为了提高训练效率和模型可扩展性，他们采用了多阶段训练，主要包括三步：

先用text-to-image(T2I) 任务作为warm-up训练，然后进行text-to-image和text-to-video的联合训练；

从低清到高清；

优化数据集和训练方案。

值得一提的是，在视频生成的评估中，常常是人类亲自来评估。但是MovieGen有一个视频验证集，能够和人类的评估标准很好地耦合。

此外，在预训练之后，MovieGen进行了进一步的微调，其使用的小而精的微调视频数据集全部由人工挑选产生，所以这一步可以理解为supervised finetuning（SFT）。

为了得到有良好的动作、真实感、美感、概念多元、高质量文字说明的高质量视频数据集，他们先在候选视频中刷掉了不符合基本要求和包含小物体的视频。接下来，剩余的百万条视频存在类别不均匀的问题，他们通过人工挑选出每种概念风格中的种子视频进行聚类、平均抽样。最后对剩余的数据集进行人工的筛选，并进行人工的文字说明（captioning），得到高质量的微调数据集。

微调视频数据集中，视频的长度在10-16s的范围内，有50%的视频长度为16s，这和先前长度分布在4-16s区间的普通数据集有一定差别，微调视频数据集主要针对高清、动作大、时间长、语义信息丰富的视频生成而设计。

最后，MovieGen将微调中不同超参数下训练出的多个checkpoint模型参数进行了平均，得到了最终的模型，这和Llama3使用了类似的方法，将各有所长的模型进行了融合。

HunyuanVideo在sft部分做的工作也和MovieGen相似，但是在captioning这一部分有较大的区别，我们进行具体的分析。混元使用了json文件来进行captioning，json文件中包括：

1) 简短描述：捕捉场景的主要内容。

2) 详尽描述：详细描述场景内容，尤其包括场景转换和与视觉内容相结合的镜头运动，例如镜头跟随某个主体。

3) 背景：描述主体所处的环境。

4) 风格：描述视频的风格，例如纪录片、电影、写实或科幻。

5) 镜头类型：识别突出或强调特定视觉内容的镜头类型，例如航拍、特写镜头、中景或远景。

6) 灯光：描述视频的灯光条件。

7) 氛围：传达视频的氛围，例如温馨、紧张或神秘。

这样的json格式的captioning从多个角度来对视频数据进行了captioning，使得食品能够更加贴合人类prompt需求。然而统一使用如上格式可能导致多样性的缺失，所以混元还设计了一个dropout机制来增加captioning的多样性。

# 模型和训练

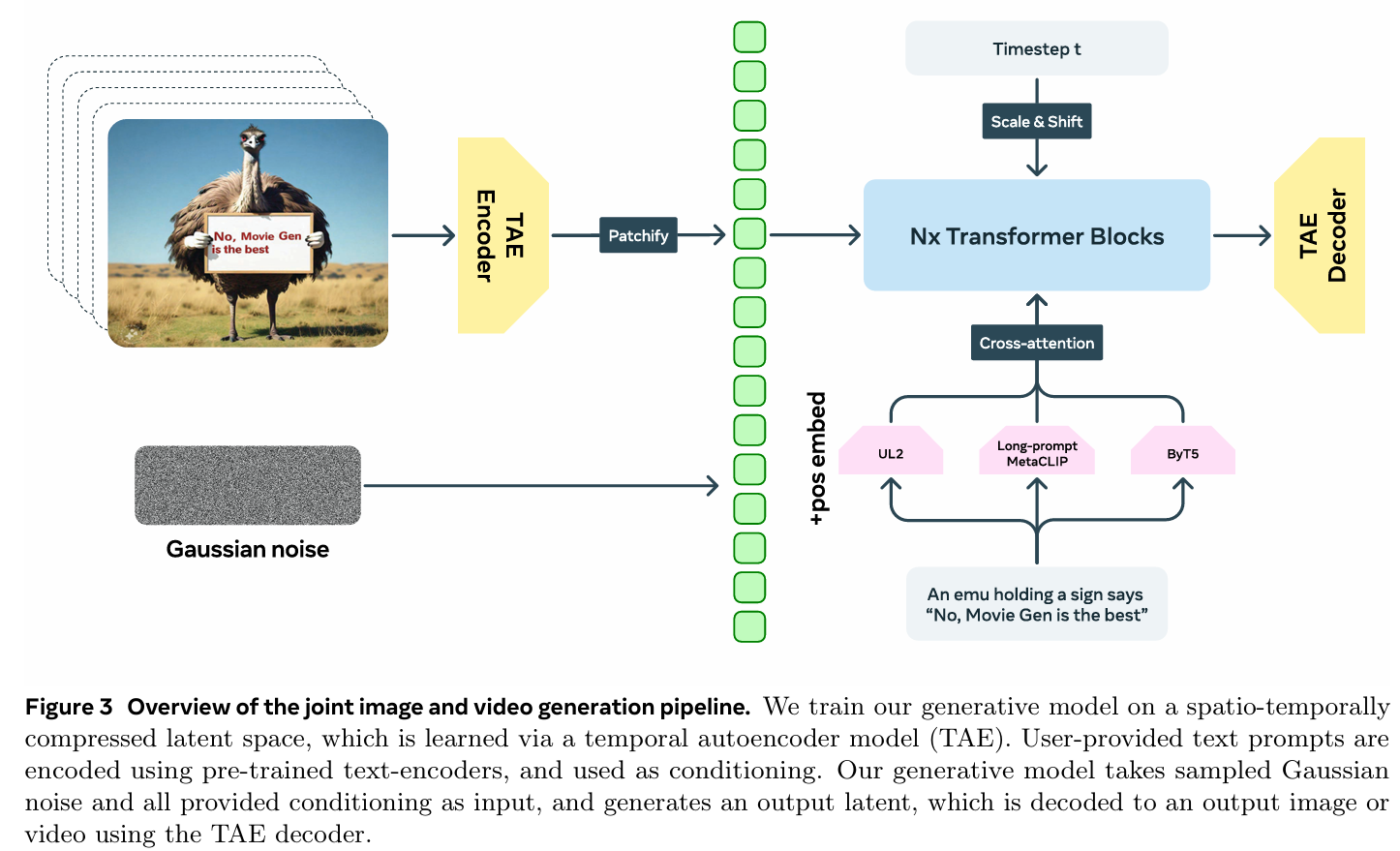
Sora的技术报告比较提纲挈领的同时并没有披露任何细节，所以先通过Sora技术报告来梳理一遍思路。

和ChatGPT相似，视频生成也是基于transformer，通过把视频处理成token后，也能获得scaling law的收益，即通过增加算力即可获得更好的训练效果。因此，Sora分两步把原始视频数据变成patch：

先把原始视频帧压缩到更低维的潜在空间。相对于图片来说，视频包含的信息要多出成千上万倍，所以通过压缩原始数据，可以减轻模型上下文长度上的压力。Sora训练了一个神经网络来完成这一步，同时他们还训练了一个decoder模型，可以把压缩后的潜在空间还原到原视频。

再进一步拆解成时空序列的patches。这一步Sora、MovieGen、HunyuanVideo都很相似，在MovieGen部分进行分析。

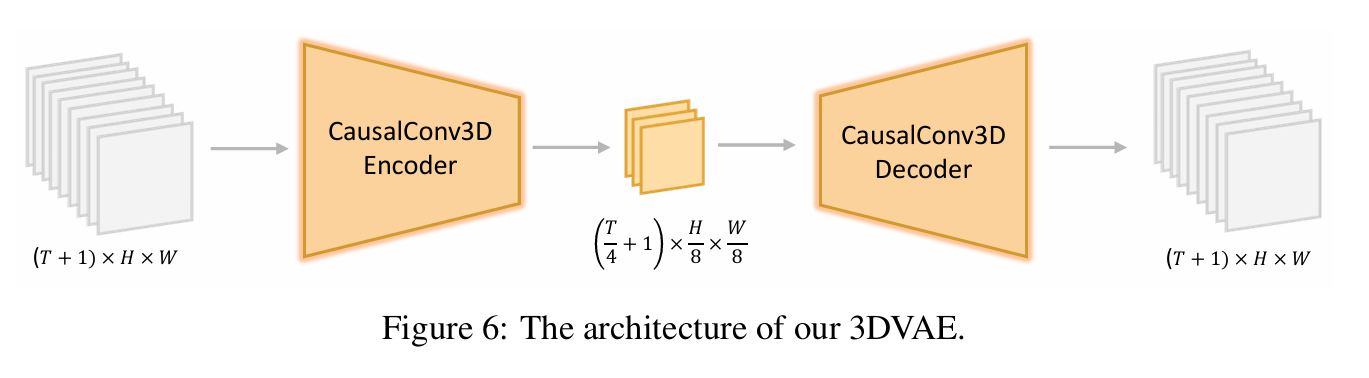
在训练阶段，Sora使用的是扩散模型，输入噪声较大的视频信号，输出则是尽可能对原信号进行还原。



## 潜在空间压缩

如上图所示的MovieGen模型架构，视频在输入之前要先经过Temporal Autoencoder(TAE) Encoder的压缩。这里的TAE Encoder和Sora的视频压缩是一回事，它基于变分自编码器[7]，将形状为T’×3×H’×W’ 的输入像素空间视频V压缩为形状为T×C×H×W 的连续值潜在空间X。在MovieGen的实现中，他们将输入在每个时空维度上压缩 8 倍，即T’/T = H’/H = W’/W = 8，实现了512倍的压缩。

具体地，TAE采用了预训练的图像自编码器架构，并通过添加时间参数对其进行“膨胀”：在每个二维空间卷积后添加一个一维时间卷积，并在每个空间注意力后添加一个一维时间注意力。



相比之下，如上图所示，HunyuanVideo训练的则是3DVAE模型。为了同时处理视频和图像，他们采用了CausalConv3D，对于形状为 (T+1)×3×H×W的视频，3DVAE将其压缩为形状为 (T/ct + 1)×C×(H/cs) × (W/cs) 的潜在空间。

MovieGen的2+1D方法和HunyuanVideo的3D方法在评估中表现不相上下，性能上没有明显差异，但2+1D的效率要比3D高出不少，不过混元的生成效果也是很不错的，这里的evaluation metrics也只是起到参考作用。

## 视频生成的主干架构

不论是Sora、MovieGen还是HunyuanVideo，他们都是先对视频进行压缩，然后使用3D卷积层将压缩后的张量展平成一维的patch序列，以便作为transformer的输入。

MovieGen使用了可学习位置嵌入，以便将任意大小、宽高比和视频长度的输入传入 Transformer，这是基于先前提出的DIT[8]。他们紧密遵循Llama3[9]架构中使用的Transformer模块构建了Transformer主干，并像前人工作一样使用了RMNSorm和SwiGLU。针对使用流匹配的视频生成，MovieGen对LLaMa3 Transformer模块进行了三处更改：

1. 在每个Transformer模块的自注意力模块和前馈网络（FFN）之间添加了一个交叉注意力模块。

2. 添加自适应层范数块，将时间步长 t 合并到 Transformer 中。

3. 使用完全双向注意力机制，而非语言建模中使用的因果注意力机制。

MovieGen了保持主干网络的设计简单，与LLaMa3类似，这种设计有利于扩展模型大小和训练。并且经过他们的实证研究，他们简单的架构设计性能与先前研究中使用的专用模块相当甚至更好，同时在不同超参数下的训练中更加稳定。

# 未来趋势与挑战

随着 Sora、MovieGen、HunyuanVideo 等视频生成模型的快速发展，视频生成正迈向更加真实、高分辨率、长时序、可控性的新时代。

未来，多模态融合与精准控制将成为主流方向。Sora 展示了文本驱动长视频生成的潜力，而 MovieGen 和 HunyuanVideo 则强调结构化先验（如脚本、动作轨迹、场景分镜）对生成质量的提升。未来的模型将更注重对文本、音频、动作、视觉布局等多模态信息的联合建模，以实现更精细的场景控制和人物行为规划。

同时，长视频生成能力是关键的技术瓶颈。当前模型仍难以生成分钟级高一致性视频，尤其在角色身份保持、空间逻辑统一、叙事连贯性等方面仍存在挑战。如何设计高效的长时序建模结构、跨帧注意机制与记忆模块，将成为突破口。

此外，评估体系、伦理与内容安全问题也逐渐凸显。现有指标难以全面评估视频生成的真实性、可控性与语义一致性，缺乏通用的定量评估标准，基本都需要人工评估。同时，随着生成能力增强，虚假内容泛滥、版权归属、社会误导等问题也愈发紧迫，亟需制定统一的生成视频审核标准与伦理框架。

综上所述，尽管当前视频生成模型已展现出令人惊艳的技术潜力，但真正实现产业级别的文本到视频生成系统仍面临诸多挑战。

# 参考文献

1. OpenAI. Sora: A text-to-video foundation model. https://openai.com/index/sora-system-card/, 2024.
2. A. Polyak \*et al\*., "Movie Gen: A Cast of Media Foundation Models," \*arXiv preprint arXiv:2410.13720\*, 2024. [Online]. Available: <https://arxiv.org/abs/2410.13720>
3. W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang \*et al\*., "HunyuanVideo: A Systematic Framework for Large Video Generative Models," \*arXiv preprint arXiv:2412.03603\*, 2024. [Online]. Available: <https://arxiv.org/abs/2412.03603>
4. X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey \*et al\*., "Emu: Enhancing image generation models using photogenic needles in a haystack," \*arXiv preprint arXiv:2309.15807\*, 2023. [Online]. Available: <https://arxiv.org/abs/2309.15807>
5. R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, and I. Misra, "Emu Video: Factorizing text-to-video generation by explicit image conditioning," in \*Proc. Eur. Conf. Comput. Vis. (ECCV)\*, 2024.
6. E. Pizzi, S. D. Roy, S. N. Ravindra, P. Goyal, and M. Douze, "A self-supervised descriptor for image copy detection," in \*Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)\*, 2022.
7. D. P. Kingma, "Auto-encoding variational Bayes," \*arXiv preprint arXiv:1312.6114\*, 2013. [Online]. Available: <https://arxiv.org/abs/1312.6114>
8. M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin \*et al\*., "Patch n’ Pack: NaViT, a vision transformer for any aspect ratio and resolution," in \*Advances in Neural Information Processing Systems (NeurIPS)\*, 2024.
9. A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan \*et al\*., "The Llama 3 herd of models," \*arXiv preprint arXiv:2407.21783\*, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783