通过后期训练增强视频生成中的场景转换感知

Hanwen Shen, Jiajie Lu, Yupeng Cao, Xiaonan Yang
Stevens Institute of Technology

Abstract

近年来，AI 生成视频在文本转视频任务上表现出色，尤其是在描绘单一场景的短片方面。然而，目前的模型在生成具有连贯场景转换的长视频方面存在困难，主要是因为它们无法从提示中推断出何时需要转换。大多数开源模型是在由单场景视频片段组成的数据集上训练的，这限制了它们学习和响应需要多个场景的提示的能力。场景转换意识的发展对多场景生成至关重要，因为它使模型能够通过准确检测转换来识别和分割视频成不同的片段。为了解决这个问题，我们提出了 Transition-Aware Video (TAV) 数据集，该数据集由经过预处理、包含多个场景转换的视频片段组成。我们的实验表明，在 TAV 数据集上进行后期训练可改善基于提示的场景转换理解，缩小所需场景与生成场景之间的差距，并保持图像质量。

1 引言

近年来，由于强大的生成模型如扩散模型（例如，Ho et al. (2020)，Song et al. (2021)，Rombach et al. (2022)，OpenAI (2023)）和视觉自回归模型（例如，van den Oord et al. (2016)，Kalchbrenner et al. (2017)，Chen et al. (2020)，Chen et al. (2023a)）的出现，从自然语言生成视觉内容的能力迅速提高。这些方法已成为现代文本到图像和文本到视频系统的核心，使得通过简单的提示就能获得高质量的结果，并构成了高级 T2V 模型的基础，如 Sora (Brooks et al., 2024) 和 Kling (Kuaishou Technology, 2024)。

我们观察到，现有的视频生成模型在处理单个场景的短片时表现良好，但在较长的、故事级别视频中常常难以维持质量和连贯性。像 EasyAnimate (Xu et al., 2024) 和 CogVideo (Hong et al., 2022) 这样的开源模型，通常难以识别场景转换的需要，即使在明确描述了多个不同场景时，它们仍然未能生成提示中指定的正确数量的场景。我们使用 50 个明确要求生成两个不同场景的提示来评估这些开源模型。如表 1 所示，生成的场景平均数量约为一个，这支持了我们关于这些模型处理多场景提示能力有限的结论。

一个可能的原因是广泛使用的视频文本数据集，例如 WebVid-10M (Bain et al., 2022)、Panda-70M (Chen et al., 2024b) 和 MiraData (Ju et al., 2024a)，在很大程度上由单场景片段组成（超过 90%），通常使用简单的场景分割技术提取。因此，目前的模型在训练过程中很少接触到明确的场景转换，这导致在需要场景更改的推理时出现分布外问题。鉴于当前模型的强大生成能力，我们探讨了通过后训练让它们在提示中识别场景转换是否可以增强整体性能，提高连贯性和视觉质量。

OpenSora	CogVideo	EasyAnimate
1.12	1.48	1.22

Table 1: 模型在明确指出两个场景的提示下生成的场景平均数量。

我们在这项工作中的贡献包括：

- 我们设计了 TAV 数据集，明确地指导模型通过后训练学习如何从提示中处理场景过渡。TAV 数据集由带有场景过渡的 10 秒视频片段及其相应的逐场景描述对组成。这些片段是从 Panda-70M 数据集中提取的，并且对于每个片段，使用大语言模型（LLM）为每个单独的场景生成单独的描述。

- 我们进行了一项实验，比较原始 OpenSora 模型和在 TAV 数据集上后训练的 OpenSora 模型生成的场景数量，使用相同的一组提示。结果表明，用 TAV 数据集进行后训练增加了场景的平均数量，表明对提示中指定的场景转换要求的理解有所改善。值得注意的是，图像质量不受影响，可通过 VBench (Huang et al., 2023) 测量。

2 相关工作

最近的研究越来越关注生成长篇、故事驱动的视频。早期的方法利用 GAN 和 VAE 来建模视频分布，而像 VideoGPT (Yan et al., 2021) 和 TATS (Ge et al., 2022) 这样的模型引入了离散潜在空间和基于 transformer 的架构以改善时间一致性。基于 transformer 的方法如 Phenaki (Villegas et al., 2022) 通过生成受文本输入条件制约的 token 序列进一步延长视频长度。最近，扩散模型成为强大的框架。像 LEO (Wang et al., 2023b) 和 LVDM (He et al., 2022) 这样的方法利用分层或潜在运动空间来合成具有增强连续性的长视频。NUWA-XL (Yin et al., 2023) 和 GAIA-1 (Hu et al., 2023) 采用结构化扩散或世界模型方法，而 FreeNoise (Qiu et al., 2023) 和 Gen-L-Video (Wang et al., 2023a) 通过聚合噪声采样或重叠段来扩展生成。StreamingT2V (Henschel et al., 2024) 提出了一种自回归框架，并通过记忆机制来维持外观的一致性。

过渡生成。 场景转换对于讲故事至关重要，使时间、空间或视角的顺畅转换成为可能。传统技术如淡入淡出、溶解、擦除和剪辑通常通过预定义模式实现，而变形方法（Wolberg (1998) , Shechtman et al. (2010) ）则通过估算像素级对应关系来实现更平滑的转换。生成方法如潜在空间插值 (Van Den Oord et al., 2017) 已被用于建模语义转换，在风格迁移 (Chen et al., 2018) 和物体变形（Sauer et al. (2022) , Kang et al. (2023) ）中应用。最近的进展探索了用于生成场景转换的数据驱动方法。Seine (Chen et al., 2023b) 引入了一种专注于转换和预测的短到长视频扩散模型。Loong (Wang et al., 2024) 利用自回归语言模型来生成分钟级多场景视频，而 VideoDirectorGPT (Lin et al., 2023) 则结合了大型语言模型引导的规划以确保多个场景之间的一致性。

自然地，整合专门设计用于提高一致性和捕捉场景转换的模块对于增强长视频生成是至关重要的。从我们的角度来看，评估训练数据集的质量并识别最适合和高效的数据用于此任务也应该是首要任务。据我们所知，这方面收到的关注有限，我们提出的 TAV 数据集旨在强调其重要性。

数据集。 公共视频–文本数据集可以大致按规模和重点进行分类。网页规模语料库——MiraData（33 万长视频片段）(Ju et al., 2024b) ，HD- VILA 100M (Xue et al., 2022) ，以及自动生成字幕的数据集，如 Panda-70M (Chen et al., 2024a) 和 InternVid (Wang et al., 2023c) ——提供了数亿对帧，用于支持分钟级扩散/Transformer 训练。以 WebVid-10M (Bain et al., 2022) 和 HowTo100M (Miech et al., 2019) 为代表的一般短视频字幕集主导了文本到视频的预训练，而像 Kinetics-700 (Carreira et al., 2019) 和 Moments-in-Time (Monfort et al., 2019) 这样的动作标签数据集则强调剪辑级语义。最后，一系列领域特定的基准仍然是评估和特定任务不可或缺的：经典识别/字幕语料库（UCF101 (Soomro et al., 2012) ， MSR-VTT (Xu et al., 2016) , ActivityNet-Captions (Krishna et al., 2017) ， YouCook2 (Zhou et al., 2018) ）；以自我为中心的 Ego4D (Grauman et al., 2022) ； 以面孔为中心的 CelebV-Text (Gu et al., 2023) ； 机器人 BAIR (Finn et al., 2017) ； 合成的 Moving MNIST (Srivastava et al., 2015) ； 以及面向插值的 Vimeo-90K (Xue et al., 2019) 。这些资源共同涵盖从数千到数亿的视频，支持着现代生成模型在训练、微调和评估中的应用。

3 方法

在本节中，我们介绍准备 TAV 数据集的流程。

数据来源。 我们首先从 Panda-70M 数据集的验证集中抽取了 500 个视频样本，该数据集中总共包含 2,000 个视频。该样本经过精心构建，以确保其类别分布与整个数据集的类别分布尽可能接近，从而有效地作为总体的代表性子集。在训练后的阶段，样本被分为 480 个视频用于训练，50 个用于验证，50 个用于测试。

我们修改了 PySceneDetect 中的方法。令 $L(i,j)$ 和 $R(i,j)$ 表示两个图像帧中同一通道在位置 (i,j) 的像素值，N 为像素总数。我们定义平均像素差为：

$$D(L,R) = \frac{1}{N}\sum_{i,j} |L(i,j) - R(i,j)|.$$

然后我们计算连续帧之间每个 HSV 通道的平均像素差，并将整体帧变化值定义为

$$\begin{aligned} V_t =& w_H \cdot D(H_t, H_{t-1}) \\ &+ w_S \cdot D(S_t, S_{t-1}) + w_V \cdot D(V_t, V_{t-1}). \end{aligned}$$

这里 w_H, w_S, w_V 是用户分配给每个通道的权重。如果

$$V_t > \text{threshold}.$$

则检测到场景切换。

我们将上述现场转换检测方法应用于先前选择的 500 视频样本。对于每个视频，我们仅保留检测到的第一个场景切换，并提取一个围绕该切换点的 10 秒剪辑（结合切换点前后的 5 秒），从而获得包含清晰场景切换的片段。如果切换点的任一侧不包含完整的 5 秒视频素材，则包括尽可能多的可用部分。

视频数据字幕。 在获取包含两个不同场景的 10 秒短片后，我们使用 BLIP 为每个场景生成单独的文本描述。然后将这些描述合并成一个明确指示场景转换的提示。例如：{ 之前的场景：超人正在城市上空飞翔；下一个场景：他看到蝙蝠侠正在屋顶上与小丑搏斗 }。TAV 数据集由以这种方式构建的 500 视频-提示对组成。

4 实验

为了一致性和简洁性，我们提请读者参考附录 A-C 查阅实现细节，包括代码和示例帧条。

我们在视频到文本生成设置下使用 TAV 数据集微调

实现。 OpenSora-Plan v1.3.1 (Lin et al., 2024) 模型。训练是在 DeepSpeed Zero Stage 2 优化的单个进程中进行的。我们使用 google/mt5-xxl 文本编码器，并采用从 OpenSora-Plan v1.3.0 预训练的 WFVAEModel（D8_4x8x8）作为视频自动编码器。模型处理 33 帧的视频剪辑，分辨率为 256 × 256，采样率为 1，帧率为 8 FPS。使用单个 H200 GPU，每个训练周期大约完成 2 小时。

关键的超参数包括批量大小为 1，总共进行 100 步训练，学习率为 1×10^{-5}，并使用常量调度器，以及 bf16 混合精度训练。我们使用指数滑动平均（EMA），其衰减率为 0.9999，从第 0 步开始。开启梯度检查点，并从最新的检查点恢复训练。其他策略包括稀疏 1D 注意（sparse_n = 4），时间和空间插值尺度设为 1.0，以及指导尺度为 0.1。模型使用 SNR 加权损失（snr_gamma = 5.0），并采用 v_prediction 类型进行扩散。

为了评估模型的性能，我们构建了三个评估组，每个组使用不同版本的提示词来测试训练后的效果。

- 组 A。该组使用由单句构成的提示，不指示任何场景转换（例如，{ 超人飞越建筑 }）。这用于证明模型也能够处理单一场景的生成，突出其在多场景转换之外的多功能性。
- 组 B。该组使用包含两句话的提示，这些句子暗示但未明确指出场景转换。例如：{ 超人在建筑物上空飞过，然后看到蝙蝠侠正在屋顶上与小丑战斗 }。
- 第 C 组。该组使用明确指示场景转换的提示。例如：{ 上一个场景：超人飞过建筑物；下一个场景：超人在屋顶上看到蝙蝠侠与小丑搏斗 }。

我们将提示（最初是从 TAV 数据集中 50 测试视频的文本描述中生成的）修订为上述三组。这些提示随后应用于基线和后训练模型，以评估场景转换的平均数量和整体图像质量。我们在表 2 中展示了结果。

5 结果与分析

如表 2 所示，后训练后场景的平均数量明显增加。在基线模型中，尤其是对于组 B 和组 C，片段的平均数量保持在 1 左右，表明其在识别多场景生成需求方面能力有限。相反，经过后训练的模型显示了显著的改进，片段的平均数量甚至超过了 2。这些结果表明，使用 TAV 数据集进行的后训练有效地增强了模型的多场景生成能力。

此外，后期训练不会明显降低视频质量。相反，它提高了动态一致性和时间平滑性，使模型能够生成更连贯的运动和流畅的场景过渡。

随着训练的进行，我们还观察到了美学质量和成像质量的逐步提高，各项指标接近或匹配基线。这些结果表明，使用 TAV 数据集进行的训后训练在不损害视觉逼真度的情况下增强了多场景生成能力。

此外，即使后训练是在多场景数据集上进行的，该模型在仅需单一场景的提示（组 A）上仍然表现良好。如表 2 所示，经后训练的模型不仅在提示明确指示双场景结构时（组 C）表现强劲，还在仅隐含建议过渡时（组 B）表现出色。

6 结论

总之，我们的实验表明，提示设计在控制 T2V 模型生成的场景数量方面起着关键作用。此外，在 TAV 数据集上进行后训练显著增强了模型识别和满足多场景生成需求的能力，特别是在提示中明确表达这种意图时。值得注意的是，我们还观察到，尽管在显式场景转换指令的提示上进行训练，经过后训练的模型对暗示两个场景但未明确说明转换的提示表现出更好的理解和响应能力。

首先，这项研究是一个初步实验。由于计算资源有限，我们仅评估了开源的最先进的 T2V 模型，并仅使用了 Panda70M 数据集中的一部分视频。其次，我们的实验目前仅专注于多场景视频片段。更全面的评估应包括单场景和多场景片段的混合。第三，尚不清楚在 T2V 设置中哪个场景检测算法表现最好。在我们的实验中，阈值配置是基于我们的经验启发性确定的。

我们的数据来自 Panda-70M，这是一个开源数据集。Panda-70M 数据集由 Snap Inc. 提供，

Table 2: 基线模型和后训练模型在不同训练轮次下的评估指标。

	group	epoch	average segments	aesthetic quality	overall consistency	dynamic degrees	imaging quality
Baseline	A	-	1.180	0.510	0.045	0.203	0.652
Baseline	B	-	1.060	0.551	0.042	0.038	0.648
Baseline	C	-	1.120	0.517	0.049	0.089	0.643
Post-trained	A	16	1.840	0.401	0.060	0.783	0.575
Post-trained	B	16	1.800	0.405	0.062	0.789	0.592
Post-trained	C	16	1.740	0.395	0.062	0.816	0.584
Post-trained	A	24	2.380	0.436	0.052	0.538	0.647
Post-trained	B	24	2.700	0.419	0.054	0.526	0.630
Post-trained	C	24	2.900	0.429	0.060	0.517	0.599
Post-trained	A	36	2.300	0.430	0.053	0.643	0.608
Post-trained	B	36	2.520	0.425	0.054	0.643	0.622
Post-trained	C	36	2.400	0.443	0.057	0.515	0.616

仅限于用于非商业、研究目的。在保留原始版权声明、许可条款和免责声明的前提下，允许再分发。内容是安全的，涵盖了多种视频领域，包括动物、风景、美食、运动、活动、车辆、教程、新闻和电视、游戏。提示是用英语生成的。

因为我们的主要结果是关于 T2V 模型生成的场景数量，而不是关于场景的内容，所以伦理问题的风险很小。所有提示都是由开源模型生成的。

我们非常感谢在撰写稿件过程中仅使用 AI 辅助工具进行语法纠正。研究的其他方面——包括概念化、实验设计、数据分析或结果解读——均未由 AI 生成或修改。所有实质性内容和结论均由作者独立开发。

7

实现细节 代码将在不久后上线 匿名空间 。

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2022. Frozen in time: A joint video and image encoder for end-to-end retrieval. *Preprint*, arXiv:2104.00650.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, and 1 others. 2024. Video generation models as world simulators. Technical report, OpenAI. Introduces the Sora text-to-video diffusion-transformer model.

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. In *arXiv:1907.06987*.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Image gpt. *arXiv preprint arXiv:2006.03622*.

Ting Chen, Saurabh Saxena, Geoffrey Hinton, and Ishan Misra. 2023a. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6710–6719.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, and 1 others. 2024a. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024b. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *Preprint*, arXiv:2402.19479.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023b. Seine: Short-to-long video diffusion model for generative transition and prediction. ArXiv preprint, 2023.

Xinyuan Chen, Chang Xu, Xiaokang Yang, Li Song, and Dacheng Tao. 2018. Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Transactions on Image Processing*, 28(2):546–560.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. 2017. Deep visual foresight for planning robot motion. In *Proceedings of the IEEE International Conference on Robotics and Automation*. BAIR Robot Pushing dataset.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, JiaBin Huang, and Devi

Parikh. 2022. Long video generation with timeagnostic vqgan and timesensitive transformer. In *European Conference on Computer Vision (ECCV)*.

Kristen Grauman, Andrew Westbury, Rohit Girdhar, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*.

Yufei Gu, Wenhao Huang, Xinyang Zhang, and 1 others. 2023. Celebv-text: A large-scale video-text dataset for realistic human generation. *arXiv preprint arXiv:2312.00734*.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*. Updated version v2 on 20 Mar 2023.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2024. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. Gaia1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023. Vbench: Comprehensive benchmark suite for video generative models. *Preprint*, arXiv:2311.17982.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024a. Miradata: A large-scale video dataset with long durations and structured captions. *Preprint*, arXiv:2407.06358.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024b. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*.

Nal Kalchbrenner, Aaron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. 2017. Video pixel networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1771–1779.

MinGuk Kang, Joonghyuk Shin, and Jaesik Park. 2023. Studiogan: A taxonomy and benchmark of gans for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kuaishou Technology. 2024. Kuaishou unveils proprietary video generation model ‘kling’; testing now available. Press release via PR Newswire. Introduces Kling with DiT backbone, 3D VAE and spatiotemporal attention.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, and 5 others. 2024. Open-sora plan: Open-source large video generation model. *Preprint*, arXiv:2412.00131.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*. Updated July 2024.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Mathew Monfort, Alex Andonian, Bolei Zhou, and 1 others. 2019. Moments in time dataset: One million videos for event understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

OpenAI. 2023. GPT-4 technical report. Technical Report arXiv:2303.08774, OpenAI. Technical report.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. 2023. Freenoise: Tuningfree longer video diffusion via noise rescheduling. arXiv preprint arXiv:2310.15169.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. Stylegan-xl: Scaling stylegan to large diverse

datasets. In *Proceedings of SIGGRAPH ' 22: Special Interest Group on Computer Graphics and Interactive Techniques Conference*, pages 49:1–49:10, Vancouver, BC, Canada.

Eli Shechtman, Alex RavAcha, Michal Irani, and StevenăM. Seitz. 2010. Regenerative morphing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 615–622, SanăFrancisco, CA, USA.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*. ArXiv preprint arXiv:2010.02502.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human action classes from videos in the wild. In *arXiv:1212.0402*.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. 2015. Unsupervised learning of video representations using lstms. In *Proceedings of the International Conference on Machine Learning*. Moving MNIST.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 4790–4798.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ruben Villegas, Mohammad Babaeizadeh, PieterJan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2022. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations (ICLR)*.

FuYun Wang, Wenshuo Chen, Guanglu Song, HanJia Ye, Yu Liu, and Hongsheng Li. 2023a. Genlvideo: Multitext to long video generation via temporal codenoising. arXiv preprint arXiv:2305.18264.

Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. 2023b. Leo: Generative latent image animator for human video synthesis. arXiv preprint arXiv:2305.03989.

Yi Wang, Yinan He, Yizhuo Li, and 1 others. 2023c. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. 2024. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*.

George Wolberg. 1998. Image morphing: A survey. *The Visual Computer*, 14(8-9):360–372.

Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. 2024. Easyanimate: A highperformance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. 2019. Video enhancement with task-oriented flow. In *International Journal of Computer Vision*. Vimeo-90K dataset.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vqvae and transformers. arXiv preprint arXiv:2104.10157.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. 2023. Nuwaxl: Diffusion over diffusion for extremely long video generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Luowei Zhou, Chenliang Xu, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of AAAI*.

A

附录 A：框架和时间线比较

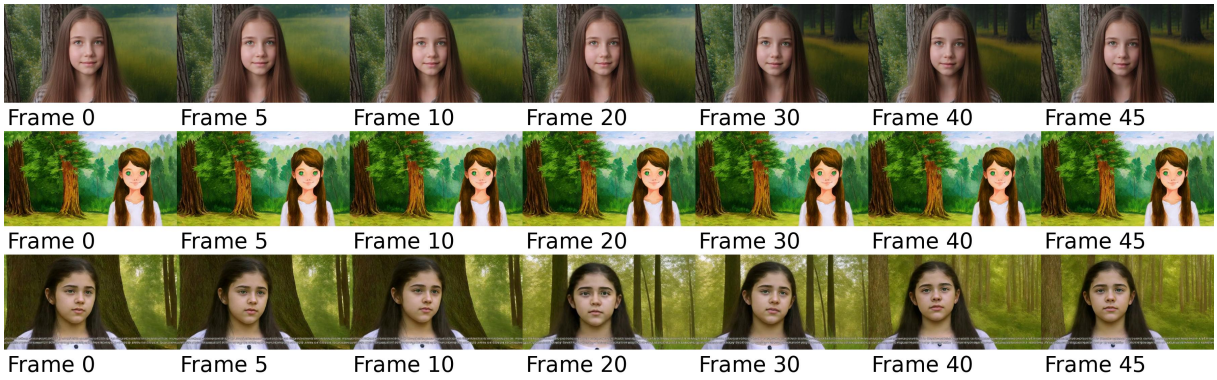

Figure 1: 三个视频生成的帧时间线比较。从上到下分别是: (i) 由 EasyAnimate 生成的输出; (ii) OpenSora-Plan 在后训练之前生成的输出; 以及 (iii) OpenSora-Plan 在 24 个周期后训练后的输出。用来生成的提示是: “前一个场景：一个长发绿眼的女孩站在一棵树前。下一个场景：一个有树木和草地的森林画作。”

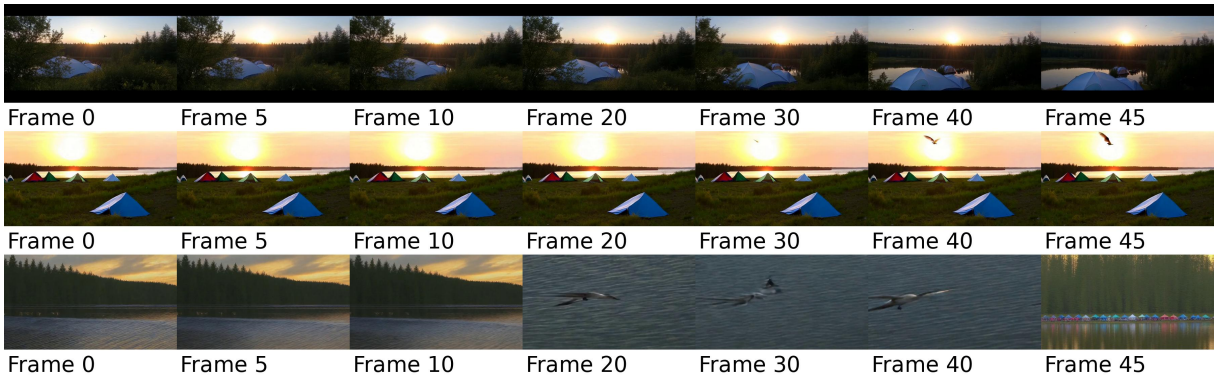

Figure 2: 三种视频生成的帧时间线比较。从上到下分别为：(i) 由 EasyAnimate 生成的输出；(ii) OpenSora-Plan 在后训练前生成的输出；以及 (iii) OpenSora-Plan 经过 24 轮后训练后的输出。用于生成的提示是：“前一幕：一群帐篷在树林里搭起；然后下一幕：夕阳下，一只鸟飞过水面。”

B

附录 B：提示示例

本附录详细列出了我们实验中使用的提示示例。提示分为三类：单场景提示、有格式的多场景提示，以及无格式的多场景提示。在总共 50 个提示中，有 42 个可以成功展现在论文中，而有 8 个由于格式或兼容性问题无法正确显示。

B.1

B.1 单场景提示

1. 一个男人和一个女人坐在海滩上的桌子旁。
2. 一组帐篷搭建在树林里。
3. 一名男子和一名女子坐在有饮料的桌子旁。
4. 一个长发绿眼的女孩站在一棵树前。
5. 一艘船在水中，靠近一座多岩石的山。
6. 一只黄黑色的鸟在蓝天中飞翔。
7. 坐着轮椅的小女孩和一个玩具。
8. 一群女性在一群人面前举着标语。
9. 顶端有钟的高塔。
10. 一个穿着西装打领带的男人正在和一个女人交谈。
11. 把那个超级英雄拿过来，给这个女孩。
12. 一个穿黑色衣服戴眼镜的女人上了新闻。
13. 一位穿比基尼的女士正在和一位男士交谈。
14. 一排放在架子上的酒瓶。
15. 一个有笔放在上面的相机的特写。
16. 一个人拿着一张带有黑白图案的白卡。
17. 一个玩偶站在床上。
18. 一个人走路时的模糊影像。
19. 一群人围绕着一棵树聚集。
20. 一位女士在新闻中坐着。
21. 一群人在街上走动。
22. 一个穿蓝色衬衫的男人站在摩托车旁边。
23. 一个人正在把一袋食物放进一个箱子里。
24. 一个人在栅栏附近的雪地里行走。
25. 一个白色盘子，上面写着“简讯”。
26. 一个戴帽子和棒球帽的男人。

27. 白色微波炉。
28. 桌子上的一个白色锅和一把银色勺子。
29. 一堆书在桌子上。
30. Adobe 中的 Adobe 文件。
31. 桌子上有几碗食物和一个碗的食物。
32. 一个装满食物的碗放在桌子上。
33. 一堆塑料袋放在桌子上。
34. 两个玩偶坐在医院的病床上。
35. 密歇根州底特律郊区的一条被淹没的街道。
36. 一只小白鼠正坐在地板上。
37. 一只猫正坐在地板上，旁边有一瓶液体。
38. 一个棒球运动员正在被裁判击中。
39. 一个卡通人物抱着一只白猫。
40. 一只猫正坐在地板上，旁边有一瓶液体。
41. 一个有标志的被雪覆盖的停车场。
42. 凤凰城，亚利桑那州的一条被水淹没的街道。

B.2

A.2 多场景提示的格式

示例格式：前一幕：……；然后下一幕：……

1. 前一场景：一男一女坐在海滩上的一张桌子旁；下一场景：一名女子坐在桌旁，桌上有一杯饮料
2. 前一场景：一群帐篷搭建在树林中；然后下一场景：一只鸟在日落时分飞过水面
3. 前一幕：一男一女坐在桌旁喝饮料；然后下一幕：一个穿比基尼的女人站在沙滩上
4. 前一个场景：一个长发和绿色眼睛的女孩站在一棵树前；下一个场景：一幅森林的画，画中有树木和草地
5. 前一幕：一艘船在水中靠近一座岩石山；然后下一幕：一位女士坐在桌子旁喝饮料
6. 前一幕：一只黄黑相间的鸟在蓝天中飞翔；接着下一幕：黄昏的女孩们
7. 前一个场景：一个坐在轮椅上的小女孩和一个玩具；然后下一个场景：一个坐在椅子上的玩偶旁边有一个盒子
8. 前一场景：一群妇女在一群人前举着标语；然后下一场景：一男一女站在麦克风前
9. 前一个场景：一个顶端有钟的高塔；然后下一个场景：一个男人正在把他的选票放入票箱
10. 前一幕：一个穿西装打领带的男人正在和一个女人交谈；然后下一幕：一个穿西装打领带的男人正在和另一个穿西装的男人交谈

11. 前一个场景：那个超级英雄由那个女孩得到；然后下一个场景：文件文件为你提供 png 文件为你提供小马宝莉

12. 之前场景：一个穿黑色连衣裙和眼镜的女人出现在新闻中；然后下一个场景：一个女人坐在电视屏幕前的沙发上

13. 前一幕：一名穿着比基尼的女性正在与一名男性交谈；然后下一幕：一名男性和一名女性正坐在桌边，桌上有饮料

14. 前一个场景：一排酒瓶放在架子上；下一个场景：一个男人站在吧台前

15. 前一幕：特写镜头中的相机上有支笔；下一幕：一名男子站在摩托车前

16. 前一幕：一个人拿着一张带有黑白图案的白卡；然后下一幕：一个男人拿着手机

17. 前一个场景：一个洋娃娃站在床上；然后下一个场景：一个小女孩正在放一个礼物盒

18. 上一场景：一个人走动的模糊身影；然后下一个场景：一个紫色花瓶，瓶中有一朵白色的花

19. 前一幕: 一群人围着一棵树聚集；然后下一幕: 一只猫站在黑暗中

20. 前一场景：一名女子坐在新闻中；然后下一场景：两个女子坐在沙发上互相交谈

21. 前一个场景：一群人在街上走动；接下来的场景：一个穿蓝色夹克的女人在街上走

22. 前一场景：一个穿蓝色衬衫的男人站在摩托车旁；然后是下一个场景：一个手机的特写

23. 前一幕：一个人正在把一袋食物放进一个盒子里；接下来一幕：一个人正在把食物放进一个容器里

24. 前一场景：一个人在篱笆附近的雪中行走；然后变为下一场景：黑色背景上的白色和红色的花

25. 前一个场景：一个白色盘子，上面写着新闻简报；然后下一个场景：一个女人站在砖墙前

26. 前一场景：一个戴着帽子和棒球帽的男人；然后下一场景：警察调查一名在河里的车后座中枪的男人

27. 前一场景：一个白色微波炉；然后下一场景：一个带勺子和杯子的白色碗

28. 上一个场景：桌子上的一个白色锅和一个银色汤匙；下一个场景：一个白色歪斜的锅

29. 前一场景：桌子上有一堆书；然后下一场景：桌子上有一堆食品盒

30. 前一个场景：Adobe 中的 Adobe 文件；然后下一个场景：绿色背景的电脑屏幕

31. 上一个场景：桌子上有几碗食物和一碗食物；然后下一个场景：制作蛋糕的原料

32. 前一个场景：一个装满食物的碗放在桌子上；下一个场景：一个白色的杯子里放着一把勺子

33. 前一场景：一堆塑料袋放在桌子上；然后下一场景：一堆塑料袋

34. 前一个场景：两个玩偶坐在医院病床上；然后下一个场景：两个玩偶坐在椅子上

35. 前一个场景：密歇根州底特律郊区一条被淹的街道；接下来的场景：一只狗站在被淹的街道中央

36. 前一个场景：一只小白鼠坐在地板上；接下来的场景：一只小狗坐在地板上

37. 前一幕：一只猫坐在地板上，旁边有一瓶液体；然后下一幕：一只小白鼠

38. 前一幕：一名棒球运动员被裁判击中；然后下一幕：一名棒球运动员正准备接球
39. 前一幕：一个卡通人物抱着一只白猫；接下来一幕：一个卡通人物处于蓝色背景中
40. 前一个场景：一只猫坐在地板上，旁边有一瓶液体；然后下一个场景：一只猫坐在地板上，旁边有一瓶酱汁
41. 前一个场景：一个覆盖着雪的停车场和一个标志；下一个场景：一个黑色背景上的白色和红色花朵
42. 前一个场景：亚利桑那州凤凰城一条被淹的街道；然后下一个场景：一堵涂满涂鸦的墙上贴着警戒线

B.3

B.3 多场景提示没有格式

1. 一名男性和一名女性坐在海滩上的一张桌子旁；一名女性与一杯饮料坐在一张桌子旁。
2. 一群帐篷搭建在树林中；一只鸟在日落时分飞过水面。
3. 一位男士和女士坐在桌子旁喝饮料；一位穿比基尼的女士正站在海滩上。
4. 一个长发碧眼的女孩站在一棵树前；这是一幅描绘森林中树木和草地的画作。
5. 一艘船在靠近岩石山的水中；一个女人坐在桌子旁，手边有一杯饮品。
6. 一只黄黑相间的鸟在蓝天中飞翔；黄昏时刻的女孩们。
7. 一个坐在轮椅上的小女孩和一个玩具；一个坐在椅子上的洋娃娃，旁边有一个箱子。
8. 一群女性在一群人前举着标语；一名男子和一名女子站在麦克风前。
9. 一个顶部有钟的高塔；一个男人正在将他的选票放入票箱。
10. 一位穿着西装和领带的男人在和一位女人交谈；一位穿着西装和领带的男人在和另一位穿着西装的男人交谈。
11. 抓到那个女超级英雄；为你准备好的文件 png 文件，为你的小马。
12. 一位穿黑色连衣裙并戴眼镜的女人上了新闻；一位坐在沙发上，面前有一个电视屏幕的女人。
13. 一个穿比基尼的女人在和一个男人讲话；一个男人和女人坐在一张桌子旁喝饮料。
14. 一排酒瓶在架子上；一个男人正站在酒吧。
15. 一台相机的特写，上面有一支笔；一个男人站在摩托车前面。
16. 一个人拿着一张带有黑白花纹的白色卡片；一个男人正在拿着手机。
17. 一个娃娃站在床上；一个小女孩正在放一个礼物盒。
18. 一个人走动的模糊；一个紫色的花瓶上插着一朵白色的花。
19. 一群人聚集在树周围；一只猫站在黑暗中。
20. 一个女人正在新闻中坐着；两个女人坐在沙发上互相交谈。
21. 一群人在街上走动；一个穿着蓝色夹克的女人在街上行走。
22. 一个穿蓝色衬衫的男人站在摩托车旁边；特写镜头是一个手机。

23. 一个人正在把一袋食物放进一个箱子里；一个人正在把食物放进一个容器里。

24. 一个人在雪地里靠近围栏行走；一个有白色和红色花朵的黑色背景。

25. 一个白色盘子上写着"news brief" 的字样；一名女子站在砖墙前。

26. 一个戴帽子和棒球帽的男人；警察调查了一名被枪击倒在河中车辆内的男子。

27. 一个白色微波炉；一个带勺子和杯子的白色碗。

28. 一只白色的锅和一只银色的勺子在桌子上；一个白色的歪曲的锅。

29. 桌子上一堆书；桌子上有一堆食品盒。

30. Adobe 中的文件；一个具有绿色背景的电脑屏幕。

31. 一张摆满食物碗和一个食物碗的桌子；做蛋糕的材料。

32. 一个装满食物的碗放在桌子上；一个里面有勺子的白色杯子。

33. 一堆塑料袋放在桌子上；一堆塑料袋。

34. 两个洋娃娃坐在病床上；两个洋娃娃坐在椅子上。

35. 密歇根州底特律郊区的一条被淹的街道；一只狗站在被淹的街道中间。

36. 一只小白鼠坐在地板上；一只小狗坐在地板上。

37. 一只猫坐在地板上，旁边是一个液体瓶子；一个小白鼠。

38. 一个棒球运动员正被裁判击中；一个棒球运动员正要接球。

39. 一个握着白猫的卡通角色；一个蓝色背景的卡通角色。

40. 一只猫正坐在地板上，旁边有一瓶液体；一只猫正坐在地板上，旁边有一瓶酱。

41. 一个被雪覆盖的停车场和一个标志；一个黑色背景上的白色和红色花朵。

42. 亚利桑那州菲尼克斯的一条被淹的街道；一个警戒线贴在被涂鸦覆盖的墙上。

A

附录 C：视频转场剪辑提取代码

Python Code for Scene Transition Detection and Clip Extraction:

```
import json
import cv2
import numpy as np
import pandas as pd
import ffmpeg
from scenedetect import detect, ContentDetector
from tqdm import tqdm
import os

 # Configuration parameters
CLIP_LENGTH = 10 # target duration in seconds
PADDING = 5 # padding before and after transition point
MIN_SCENE_LENGTH = 3
MAX_SCENE_LENGTH = 10

def detect_scenes(video_path):
  "Detect scene transitions using PySceneDetect"
  scene_list = detect(video_path, ContentDetector())
  return [scene[1].get_seconds() for scene in scene_list]

def extract_transitional_clips(video_path, scene_timestamps):
  video_name = os.path.basename(video_path).split('.')[0]
  output_clips = []
  cap = cv2.VideoCapture(video_path)
  fps = cap.get(cv2.CAP_PROP_FPS)
  total_frames = int(cap.get(cv2.CAP_PROP_FRAME_COUNT))
  video_duration = total_frames / fps
  for timestamp in scene_timestamps:
    start_time = max(0, timestamp - PADDING)
    end_time = min(video_duration, timestamp + PADDING)
    if end_time - start_time > MAX_SCENE_LENGTH:
       end_time = start_time + MAX_SCENE_LENGTH
     output_filename = f" { video_name } _ { int(start_time) } - { int(end_time) }
.mp4"
    output_path = os.path.join(OUTPUT_VIDEO_DIR, output_filename)
    ffmpeg.input(video_path, ss=start_time, to=end_time)
       .output(output_path, vcodec="libx264", acodec="aac")
       .run(overwrite_output=True, quiet=True)
    output_clips.append( {
       "file_path": output_path, "video_name": video_name,
       "start_time": start_time, "end_time": end_time,
       "duration": end_time - start_time, "transition_frame": timestamp
     } )
  cap.release()
  return output_clips
```

Figure 3: 用于检测场景转化并提取以转化为中心的固定长度视频剪辑的 Python 代码。

```
def validate_clips(clips):
  filtered_clips = []
  for clip in tqdm(clips, desc="Validating Clips"):
    cap = cv2.VideoCapture(clip["file_path"])
    prev_frame = None; transition_detected = False
    while cap.isOpened():
       ret, frame = cap.read(); if not ret: break
       if prev_frame is not None:
         diff = np.mean(cv2.absdiff(prev_frame, frame))
         if diff > 50: transition_detected = True; break
       prev_frame = frame
    cap.release()
    if transition_detected: filtered_clips.append(clip)
  return filtered_clips

def save_metadata_to_json(filtered_clips):
  output_data = [ { "file_path": c["file_path"], "text": "" } for c in filtered_clips]
  with open(OUTPUT_JSON_FILE, "w", encoding="utf-8") as f:
    json.dump(output_data, f, ensure_ascii=False, indent=4)
  print(f"Metadata saved to { OUTPUT_JSON_FILE } ")

def main():
  video_path = ".../input_videos/example.mp4"
  scene_timestamps = detect_scenes(video_path)
  video_clips = extract_transitional_clips(video_path, scene_timestamps)
  validated_clips = validate_clips(video_clips)
  save_metadata_to_json(validated_clips)

if __name__ == "__main__": main()
```

Figure 4: 用于检测场景转换并提取以转换为中心的固定长度视频片段的 Python 代码。（续）