FreeLong: Training-Free Long Video Generation with SpectralBlend Temporal Attention

Yu Lu[1], **Yuanzhi Liang**[2], **Linchao Zhu**[1], **Yi Yang**[1]*

[1]The State Key Lab of Brain-Machine Intelligence, Zhejiang University
[2]University of Technology Sydney

Abstract

Video diffusion models have made substantial progress in various video generation applications. However, training models for long video generation tasks require significant computational and data resources, posing a challenge to developing long video diffusion models. This paper investigates a straightforward and training-free approach to extend an existing short video diffusion model (*e.g.*, pre-trained on 16-frame videos) for consistent long video generation (*e.g.*, 128 frames). Our preliminary observation has found that directly applying the short video diffusion model to generate long videos can lead to severe video quality degradation. Further investigation reveals that this degradation is primarily due to the distortion of high-frequency components in long videos, characterized by a decrease in spatial high-frequency components and an increase in temporal high-frequency components. Motivated by this, we propose a novel solution named FreeLong to balance the frequency distribution of long video features during the denoising process. FreeLong blends the low-frequency components of global video features, which encapsulate the entire video sequence, with the high-frequency components of local video features that focus on shorter subsequences of frames. This approach maintains global consistency while incorporating diverse and high-quality spatiotemporal details from local videos, enhancing both the consistency and fidelity of long video generation. We evaluated FreeLong on multiple base video diffusion models and observed significant improvements. Additionally, our method supports coherent multi-prompt generation, ensuring both visual coherence and seamless transitions between scenes. *Our project page is at: https://yulu.net.cn/freelong*.

1 Introduction

Video diffusion models [1, 2, 3, 4, 5, 6, 7] trained on vast video-text datasets [8, 9] have demonstrated impressive capabilities in generating high-quality videos. Inspired by Sora [10], multiple studies [11, 12, 13] have concentrated on training these models to create longer videos using extensive, long video-text datasets [14, 15, 16, 17, 18]. However, these methods demand significant computational resources and data annotations.

A more practical approach involves adapting pre-trained short video models to generate consistent longer video sequences without retraining. Recent research [19, 20] has explored sliding window temporal attention to ensure smooth transitions between video clips in the generation of long videos. Nonetheless, these techniques often struggle to maintain global temporal consistency across extended sequences and require multiple passes of temporal attention.

In this study, we propose a simple, training-free method to adapt existing short video diffusion models (e.g., pretrained on 16 frames) for generating consistent long videos (e.g., 128 frames). Initially,

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024)

 https://doi.org/10.52202/079017-4177

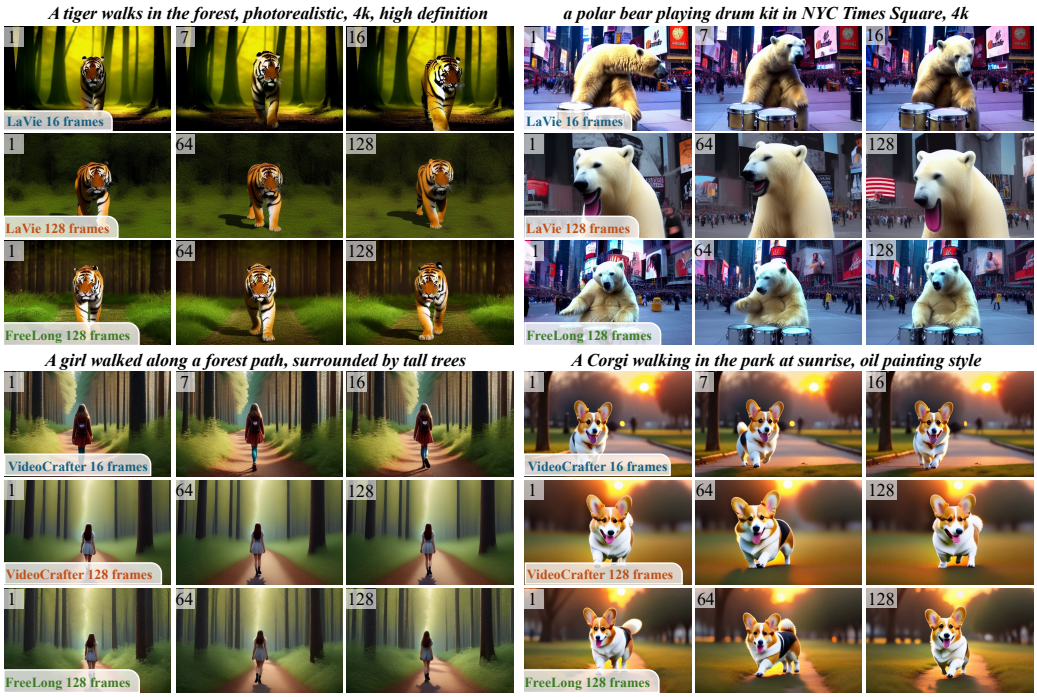


Figure 1: **Results of Short and Long Videos.** The first row of each case shows 16-frame videos generated using short video diffusion models (LaVie [1] and VideoCrafter2 [2]). Directly extending these models to longer videos, like those with 128 frames, preserves temporal consistency but lacks fine spatial-temporal details. In contrast, our proposed FreeLong adapts short video diffusion models to create consistent long videos with high fidelity.

we examine the direct application of short video diffusion models for long video generation. As depicted in Figure 1, straightforwardly using a 16-frame video diffusion model to produce 128-frame sequences yields globally consistent yet low-quality results.

To delve further into these issues, we conducted a frequency analysis of the generated long videos. As shown in Figure 2 (a), the low-frequency components remain stable as the video length increases, while the high-frequency components exhibit a noticeable decline, leading to a drop in video quality. The findings indicate that although the overall temporal structure is preserved, fine-grained details suffer notably in longer sequences. Specifically, there is a decrease in high-frequency spatial components (Figure 2 (b)) and an increase in high-frequency temporal components (Figure 2 (c)). This high-frequency distortion poses a challenge in maintaining high fidelity over extended sequences. As illustrated in the middle row of each case in Figure 1, intricate textures like forest paths or sunrises become blurred and less defined, while temporal flickering and sudden changes disrupt the video's narrative flow.

To tackle these challenges, we introduce FreeLong, a novel framework that employs SpectralBlend Temporal Attention (SpectralBlend-TA) to balance the frequency distribution of long video features in the denoising process. SpectralBlend-TA integrates global and local features via two parallel streams, enhancing the fidelity and consistency of long video generation. The global stream deals with the entire video sequence, capturing extensive dependencies and themes for narrative continuity. Meanwhile, the local stream focuses on shorter frame subsequences to retain fine details and smooth transitions, preserving high-frequency spatial and temporal information. SpectralBlend-TA combines global and local video features in the frequency domain, improving both consistency and fidelity by blending low-frequency global components with high-frequency local components. Our method is entirely training-free and allows for the easy integration of FreeLong into existing video diffusion models by adjusting the original temporal attention of video diffusion models. Comparative

https://doi.org/10.52202/079017-4177

experiments demonstrate significant improvements in temporal consistency and video fidelity when applying our method to generate long video sequences.

Our contributions can be summarized as follows: **1)** We conduct a frequency analysis on the direct application of short video models for longer video generation and identify high-frequency distortions in the longer videos. **2)** We devise a SpectralBlend Temporal Attention mechanism to merge the consistent low-frequency components of global videos with the high-fidelity high-frequency components of local videos. **3)** Our training-free approach, FreeLong, outperforms existing state-of-the-art models in both temporal consistency and video fidelity.

2 Related Work

Text-to-Video Diffusion Models: Text-to-video (T2V) generation has progressed significantly from early variational autoencoders [21, 22] and GANs [23] to advanced diffusion-based techniques [3, 4, 24, 25, 26], marking a major leap in synthesis methods. Modern video diffusion models build on pre-trained image-to-text diffusion models [27, 28, 29], incorporating temporal transformers in the diffusion UNet to capture temporal relationships. These models achieve impressive video generation results through post-training on video-text data [14, 9, 8, 1], enhancing coherence and fidelity. However, due to computational constraints and limited dataset availability, current video diffusion models are typically trained on fixed-length short videos (*e.g.*, 16 frames), limiting their ability to produce longer videos. In this paper, we propose extending these short video diffusion models to generate long and consistent videos without requiring any additional training videos.

Long-video Generation: Generating long videos is challenging due to temporal complexity, resource constraints, and the need for content consistency. Recent advancements focus on improving temporal coherence and visual quality using GAN-based [30, 31] and diffusion-based techniques [32, 33, 34, 35, 36]. For instance, Nuwa-XL [36] employs a parallel diffusion process, while StreamingT2V [11] uses an autoregressive approach with a short-long memory block to improve the consistency of long video sequences. Despite their effectiveness, these methods require substantial computational resources and large-scale datasets. Recent research has explored training-free adaptations using short video diffusion models for long video generation. Gen-L-Video [37] extends videos by merging overlapping sub-segments with a sliding-window method during denoising. FreeNoise [19] employs sliding-window temporal attention and a noise initialization strategy to maintain temporal consistency. However, these approaches focus on smooth transitions between video clips and fail to capture global consistency across long video sequences. This paper proposes FreeLong, a novel approach that blends global and local video features during the denoising process to enhance both global temporal consistency and visual quality in long video generation.

3 Observation and Analysis

When attempting to adapt short video diffusion models to generate long videos, a straightforward approach is to input a longer noise sequence into the short video models. The temporal transformer layers in the video diffusion model are not constrained by input length, making this method seemingly viable. However, our empirical study reveals significant challenges, as demonstrated in Figure 1. Generated long videos often exhibit fewer detailed textures, such as blurred forests in the background, and more irregular variations, like abrupt changes in motion. We attribute these issues to two main factors: the limitations of the temporal attention mechanism and the distortion of high-frequency components.

Attention Mechanism Analysis: The temporal attention mechanism in video diffusion models is pre-trained on fixed-length videos, which complicates its ability to generate longer videos. As shown in Figure 3 , increasing video length hinders the temporal attention's ability to accurately capture frame-to-frame relationships. For 16-frame videos, the attention maps show a diagonal pattern, indicating high correlations with adjacent frames that preserve spatial-temporal details and motion patterns. In contrast, for 128-frame videos, the less structured attention maps suggest difficulty in focusing on relevant information across distant frames, leading to missed subtle motion patterns and over-smoothed or blurred generations.

Frequency Analysis: To better understand the generation process of long videos, we analyzed the frequency components in videos of varying lengths using the Signal-to-Noise Ratio (SNR) as a

 https://doi.org/10.52202/079017-4177

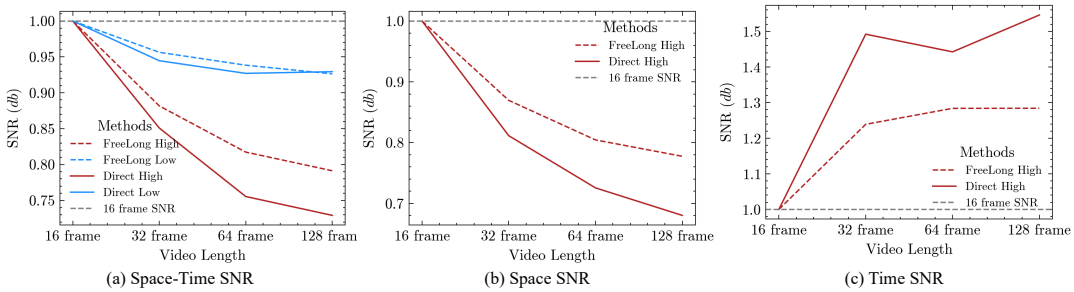


Figure 2: **Ratio of short video SNR on high/low frequency to different long videos.** Our findings reveal that: (a) When direct extend short video diffusion model to generate long videos, the SNR of high-frequency components in the space-time frequency domain degrades significantly as video length increases. (b) In the spatial frequency domain, the SNR of high-frequency components decreases even more substantially, resulting in the over-smoothing of each frame. (c) Conversely, in the temporal frequency domain, the SNR of high-frequency components increases significantly, introducing temporal flickering.

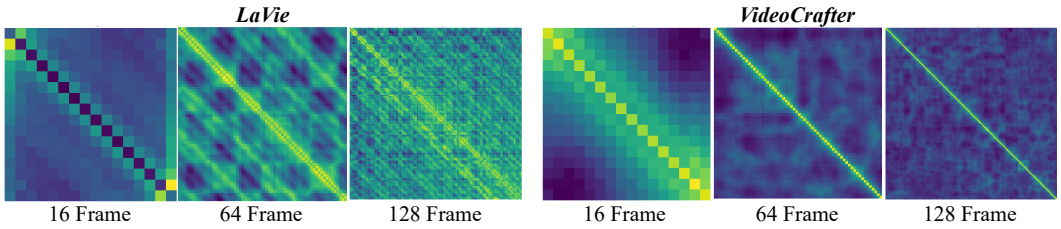


Figure 3: **Temporal Attention Visualization.** We visualize the temporal attention by average across all layers and time steps from LaVie [1] and VideoCrafter [2]. The attention maps for 16-frame videos exhibit a diagonal-like pattern, indicating a high correlation with adjacent frames, which helps preserve high-frequency details and motion patterns when generating new frames. In contrast, attention maps for longer videos are less structured, such as 128 frames, making the model struggle to identify and attend to the relevant information across distant frames. This lack of structure in the attention maps results in the distortion of high-frequency components of long videos, which results in the degradation of fine spatial-temporal details.

metric. Ideally, short video diffusion models generate 16-frame videos with high quality, and robust longer videos derived from such models should exhibit consistent SNR values across all frequency components. However, Figure 2 reveals significant differences in the SNR of high/low frequency components[2] between generated short and long videos. The SNR of low-frequency components remains relatively consistent for long videos (1.0 for 16 frames to 0.93 for 128 frames), suggesting that the model maintains overall structure and low-frequency details in extended sequences. However, the SNR of high-frequency components drops significantly for longer videos (1.0 for 16 frames to 0.73 for 128 frames), indicating a loss of fine details and increased distortion, leading to suboptimal visual fidelity.

Further investigation into the spatial and temporal frequency domains revealed two key findings: (1) In the spatial domain, the high-frequency components of long videos degrade significantly (0.68 for 128 frames), causing substantial degradation of spatial details in each frame and resulting in blurred frames. (2) In the temporal domain, the high-frequency components increase with video length (1.5 for 128-frame videos), resulting in temporal flickering and incoherent video outputs.

[2]We split the frequency components into high-frequency ($\phi \sim (0.25\pi - 1.00\pi)$) and low-frequency ($\phi \sim (0.00\pi - 0.25\pi)$) and compared the SNR of each component in long videos to the corresponding SNR in 16-frame videos.

https://doi.org/10.52202/079017-4177

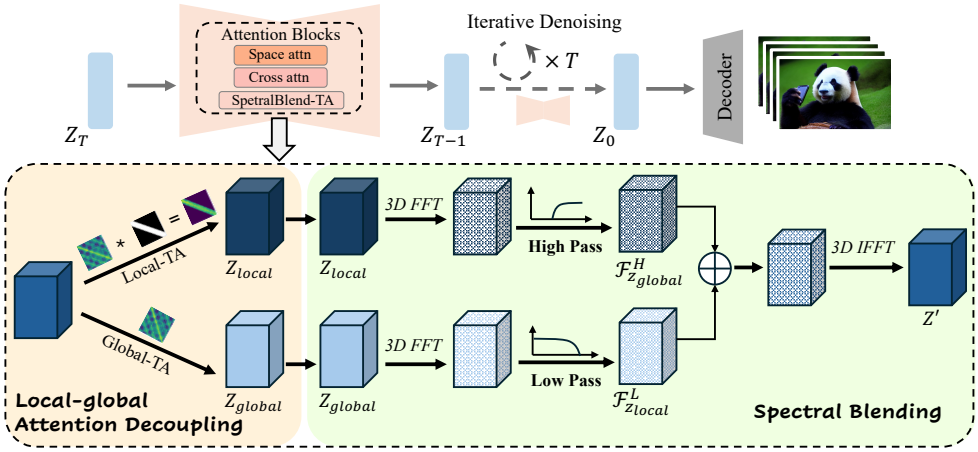


Figure 4: **Overview of FreeLong.** FreeLong facilitates consistent and high-fidelity video generation using SpectralBlend Temporal Attention (SpectralBlend-TA). SpectralBlend-TA effectively blends low-frequency global video features with high-frequency local video features through a two-step process: local-global attention decoupling and spectral blending. Local video features are obtained by masking temporal attention to concentrate on fixed-length adjacent frames, while global temporal attention encompasses all frames. During spectral blending, 3D FFT projects features into the frequency domain, where high-frequency local components and low-frequency global components are merged. The resulting blended feature, transformed back to the time domain via IFFT, is then utilized in the subsequent block for refined video generation.

4 FreeLong: Training-free Long Video Generation

Motivated by the above analysis, we propose FreeLong, a method designed to generate high-fidelity and consistent long videos using the inherent power of the diffusion model. As illustrated in Figure 4, our FreeLong uses a diffusion UNet from pre-trained short video diffusion models and introduces a SpectralBlend Temporal Attention (SpectralBlend-TA) to facilitate long video generation. The SpectralBlend-TA consists of two steps: local-global attention decoupling and spectral blending.

Local-global Attention Decoupling:

The temporal attention in short video models is optimized to model short frame sequences accurately, maintaining high-fidelity visual information. Conversely, the long-range temporal attention from short video models tends to maintain overall layout and and object consistency. Given these properties, we first decouple the local and global attention. The local attention matrix can be obtained as:

$$A_{\text{local}}(i,j) = \begin{cases} \text{Softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d}}\right) & \text{if } |i-j| \le \alpha \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where Q and K are the query and key matrices derived from the input video feature Z_{in}. The local attention A_{local} leads to each frame i only attending to frames within a window of 2α frames. Given the local attention matrix A_{local}, the local video features Z_{local} can be obtained by: $Z_{\text{local}} = A_{\text{local}} V$, where V is the value matrix derived from the input video feature Z_{in}. By restricting the temporal attention to adjacent local frames, we preserve the capabilities of short video models, thereby retaining high-fidelity visual details in local video features.

We then define the global attention matrix where each frame attends to all other frames. The global attention matrix can be computed as follows:

$$A_{\text{global}}(i,j) = \text{Softmax}\left(\frac{Q_i K_j^\top}{\sqrt{d}}\right), \tag{2}$$

Given the global attention matrix A_{global}, the global video features Z_{global} can be obtained by: $Z_{\text{global}} = A_{\text{global}} V$. The global video features process the entire video sequence, ensuring narrative continuity and coherence, while capturing long-range dependencies and overarching themes.

https://doi.org/10.52202/079017-4177

Spectral Blending: After obtaining the global and local video features, a frequency filter is used to blend the low-frequency components of the global video latent Z_{global} with the high-frequency components of the local video latent Z_{local}, resulting in a new video latent Z'. This fused latent retains the global coherence and structure provided by Z_{global}, while benefiting from the enhanced high-frequency details introduced by Z_{local}. The process is described by:

$$\mathcal{F}^{L}_{z_{global}} = \text{FFT}_{\text{3D}}(Z_{global}) \odot \mathcal{P}, \tag{3}$$

$$\mathcal{F}^{H}_{z_{local}} = \text{FFT}_{\text{3D}}(Z_{local}) \odot (1 - \mathcal{P}), \tag{4}$$

$$Z' = \text{IFFT}_{\text{3D}}(\mathcal{F}^{L}_{z_{global}} + \mathcal{F}^{H}_{z_{local}}) \tag{5}$$

where FFT_{3D} is the Fast Fourier Transformation operated on both spatial and temporal dimensions, IFFT_{3D} is the Inverse Fast Fourier Transformation that maps back the blended representation Z' from the frequency domain, and $\mathcal{P} \in \mathbb{R}^{4\times N\times h\times w}$ is the spatial-temporal Low Pass Filter (LPF), which is a tensor of the same shape as the latent. The final fused video feature Z' serves as the input to our subsequent video generation module.

The rationale behind using low-frequency components from the global video features and high-frequency components from the local video features stems from our analysis. The global features provide a stable, coherent structure, preserving the overall layout and object consistency throughout the video. This is crucial for maintaining temporal consistency in long videos. On the other hand, local features retain high-fidelity details, which are essential for capturing fine textures and intricate motion patterns that tend to degrade in long sequences. By blending these components in the frequency domain, we harness the strengths of both global consistency and local detail preservation, addressing the issues of blurred frames and temporal flickering observed in our analysis.

Recent studies [38, 39] indicate that latent diffusion models [27] generate varying levels of visual content at different stages of the denoising process: scene layout and object shapes in the early steps, and fine details in the later steps. We propose fusing global and local video features in the early τ steps of the denoising process and using local video features in the remaining steps. This fusion ensures that the overall layout and object appearance of the generated long video follow the global features, thereby maintaining temporal consistency in the generated videos.

5 Experiments

5.1 Implementation Details

Baseline Models: To evaluate the effectiveness and generalization of our proposed method, we apply FreeLong on two publicly available diffusion-based text-to-video models, LaVie [1] and VideoCrafter [2]. These models are trained on short videos with fixed length (*i.e.*, 16 frames), we extend them to produce long videos (*i.e.*, 128 frames [40]). We set $\alpha = 8$ for the local attention setting and set τ to 25. During inference, the parameters of the frequency filter for each model are kept the same for a fair comparison. Specifically, we use a Gaussian Low Pass Filter (GLPF) $\mathcal{P}_G$ with a normalized spatiotemporal stop frequency of $D_0 = 0.25$. Multi-prompt videos are generated with random noise, and FreeNoise [19] is used for single-prompt long video generation.

Test Prompts: We chose 200 prompts from VBench [41] to validate the effectiveness of our method.

Evaluation Metrics: For text-to-video generation, we employed several metrics from VBench [41] to evaluate two aspects: video consistency and video fidelity. For video consistency measurement, we use two metrics: 1). Subject consistency, computed by the DINO [42] feature similarity across frames to assess whether object appearance remains consistent throughout the whole video. 2). Background consistency, calculated by CLIP [43] feature similarity across frames. For video fidelity measurement, we use three metrics: 1). Motion smoothness, which utilizes the motion priors in the video frame interpolation model AMT [44] to evaluate the smoothness of generated motions. 2). Temporal flickering, which takes static frames and computes the mean absolute difference across frames. 3). Imaging quality, calculated using the MUSIQ [45] image quality predictor trained on the SPAQ [46] dataset.

Table 1: **Quantitative Comparison**. "Direct sampling" and "Sliding window" indicate directly sampling 128 frames and applying temporal sliding windows based on short video generation models, respectively. Compared to these methods, our FreeLong achieves consistent long video generation with high fidelity.

Methods	Sub (↑)	Back (↑)	Motion (↑)	Flicker (↑)	Imaging (↑)	Inference Time (↓)
Direct sampling	88.95	93.23	92.77	91.44	64.76	1.8s
Sliding window	85.80	92.83	95.79	94.00	66.57	2.6s
FreeNoise [19]	92.30	95.87	96.32	94.94	67.14	2.6s
Ours	**95.16**	**96.80**	**96.85**	**96.04**	**67.55**	**2.2s**

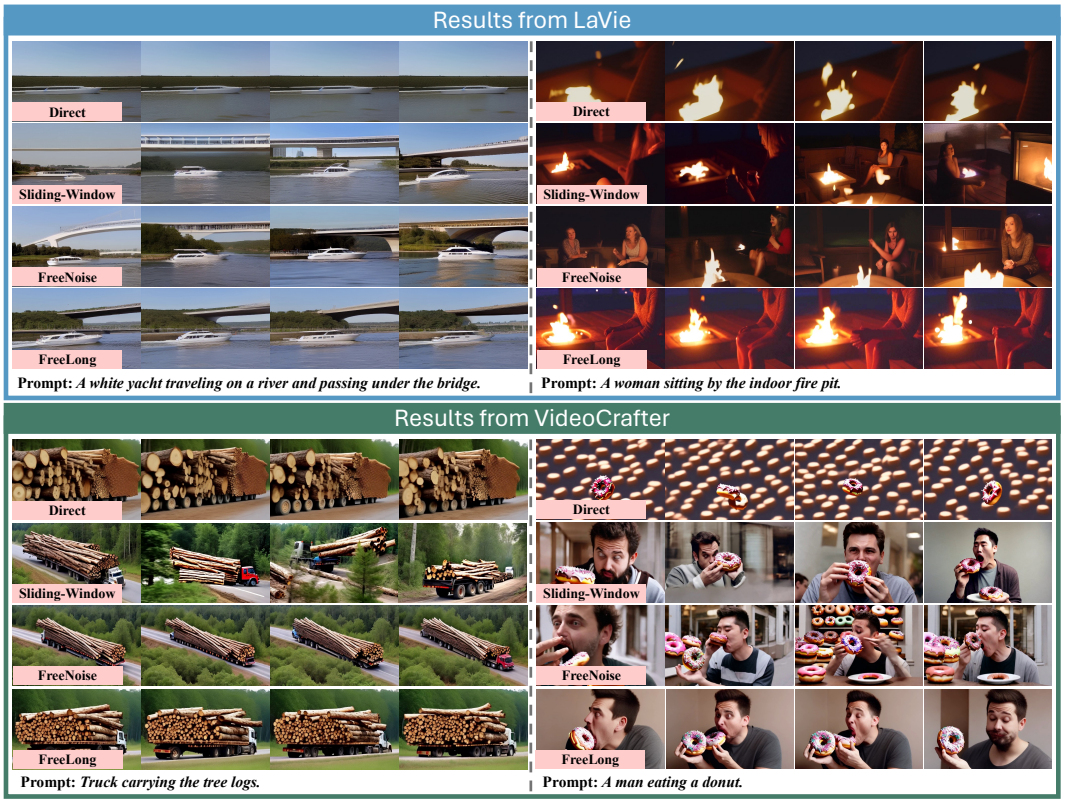


Figure 5: **Qualitative Comparison.** Results from LaVie [1] and VideoCrafter [2] are presented. Direct videos exhibit consistent frames, but they appear over-smoothed. FreeNoise and the sliding-window approach struggle to capture global consistency effectively. Our FreeLong method achieves consistent long video generation while maintaining high fidelity, preserving crucial details and textures across the entire sequence.

5.2 Quantitative Comparison

We compare our FreeLong method with other training-free approaches for long video generation using diffusion models. Our comparison includes three methods: (1) Direct sampling. It directly samples 128 frames from the short video models. (2) Sliding window. It adopts temporal sliding windows [20] to process a fixed number of frames at a time. (3) FreeNoise [19]. FreeNoise introduces repeat input noise to maintain temporal coherence across long sequences.

Table 1 presents the quantitative results. Direct generation of long videos suffers from high-frequency distortion, leading to significant quality degradation. This method results in low fidelity scores, including imaging quality, temporal flickering, and motion smoothness. The sliding-window method and FreeNoise show improved video quality thanks to the fixed effective temporal attention window

https://doi.org/10.52202/079017-4177

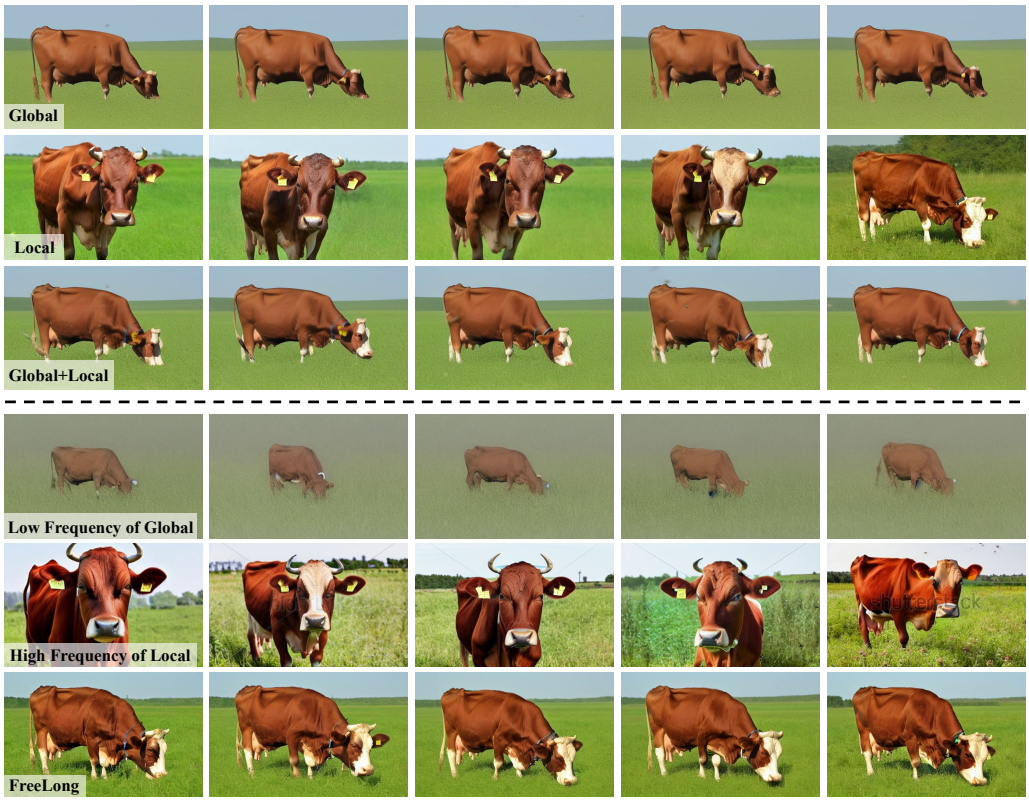
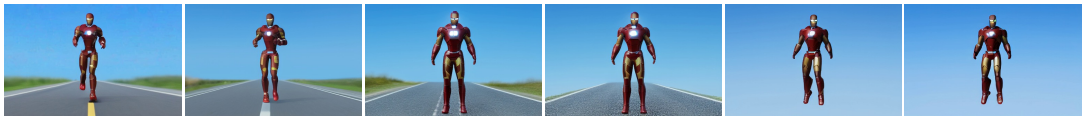


Figure 6: **Ablation Study.** Global features and low-frequency components of global features ensure consistency but degrade fidelity. Local features and high-frequency local features maintain spatial-temporal details but lack temporal consistency. Directly adding global and local features degrades fidelity. Our method achieves both high fidelity and temporal consistency.

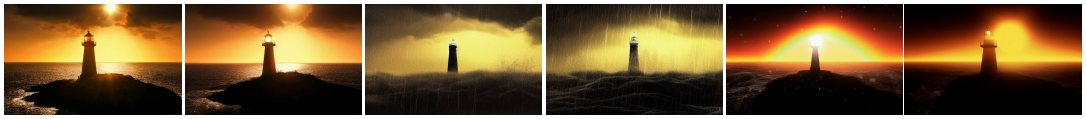
but still face challenges in maintaining consistency across long videos. Our FreeLong method achieves the highest scores across all metrics, producing consistent long videos with high fidelity. Moreover, we also examine the inference time of these methods on the NVIDIA A100. As delineated in Table 1, our approach achieves a faster speed compared to preceding methods by employing single-pass temporal attentions.

5.3 Qualitative Comparison

The synthesis results of each method are shown in Figure 5. In the first row, directly sampling 128 frames through a model trained on 16 frames will bring poor quality results due to the high-frequency distortion. For example, the yacht (left) and the girl (right) have blurred and the background is not clear. As shown in second row in Figure 5, using temporal sliding windows helps generate more vivid videos, but this approach ignores long-range visual consistency, causing the subject and background to appear significantly different across frames. FreeNoise attempts to promote global consistency by repeating and shuffling the initial noise for each frame; however, it fails to maintain long-range visual consistency and suffers from content mutations. In contrast, our method, FreeLong, explicitly enforces global constraints during the denoising process, achieving temporal consistency while preserving high fidelity across frames. Results shown in Figure 5 demonstrate that FreeLong successfully renders temporal consistent longer videos, outperforming all other methods.



Prompt1: Ironman running on the road, 4K, high resolution
Prompt2: Ironman standing on the road, 4K, high resolution
Prompt3: Ironman flying on the sky, 4K, high resolution



Prompt1: The morning sun rises, illuminating a solitary lighthouse on a rocky shore.
Prompt2: Clouds gather, and rain starts to fall, the lighthouse stand firm under storm with heavy rain and thunder.
Prompt3: The storm clears, the lighthouse stand under rainbow.
Prompt4: Night falls, and the lighthouse raise light through the darkness.

Prompt1: A musician with a guitar performs at the edge of a mountain at morning.;
Prompt2: The musician performs in a busy urban park, a crowd gathers, enchanted by the music.;
Prompt3: A serene lakeside at sunset where the musician plays alone, reflecting.;
Prompt4: Night falls, and the musician joins a lively street festival, lights and music filling the air.;
Prompt4: The journey ends on a quiet balcony overlooking the city at night, the musician performs towards the city.

Figure 7: **Results of Multi-Prompt Video Generation.** Our method ensures coherent visual continuity and motion consistency across different video segments.

5.4 Ablation Studies

To validate the effectiveness of each module in our FreeLong method—global video feature, local video feature, and our combined approach—we present the generated results by ablating each component.

As shown in the top part of Figure 6, videos generated solely from global video features maintain consistent content but suffer from severe fidelity degradation. Conversely, videos generated using only local video features preserve fidelity due to the fixed effective temporal attention window but fail to maintain temporal consistency, as evidenced by the changing color of the cow. Simply combining global and local video features results in fidelity degradation because the high-frequency components of the global video features degrade significantly.

In the bottom part of Figure 6, we show the videos generated by combining the low-frequency components from global video features with the high-frequency components from local video features. Our approach effectively combines the consistency of global videos with the high fidelity of local videos, achieving both high fidelity and temporal consistency.

5.5 Multi-Prompt Video Generation

Our method can be seamlessly extended to multi-prompt video generation by providing different prompts for each video segment. As illustrated in Figure 7, our approach ensures coherent visual continuity and motion consistency. For instance, Ironman is shown running on the road, then standing, and finally flying into the sky, all within a consistent scene and with smooth action transitions. In the second row, we demonstrate a more complex prompt sequence describing weather and scene transitions. Our method effectively models the transition from “sunrise" to “storm with heavy rain and thunder" to the final “rainbow," maintaining consistency and capturing the fine-grained details of each prompt transition.

5.6 Longer Video Generation

To examine the scalability of our FreeLong, we extend the video generation length beyond 128 frames. As depicted in Figure 8, FreeLong effectively generates videos with even longer durations, such as

https://doi.org/10.52202/079017-4177

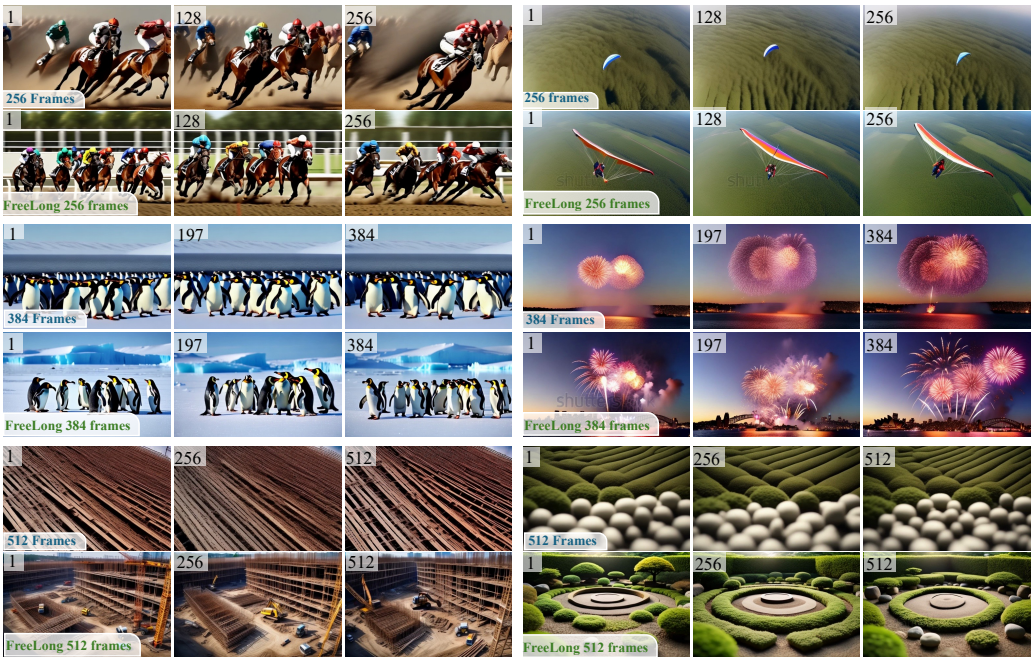


Figure 8: **Longer Video Generation.** FreeLong scales to generate videos longer than 128 frames (e.g., 512 frames), maintaining both temporal consistency and high fidelity across the entire sequence.

512 frames, while maintaining both temporal consistency and high fidelity throughout the entire sequence. This demonstrates that our method scales well with increasing video lengths, addressing the challenges associated with generating long continuous content without significant degradation in quality.

6 Conclusion

In this paper, we introduced FreeLong, a training-free method to adapt short video diffusion models for long video generation. Our research reveals that directly generating long videos from short video diffusion models results in poor quality, primarily due to high-frequency distortion. To resolve this issue, we employ the SpectralBlend Temporal Attention (SpectralBlend-TA) mechanism, which blends low-frequency global features with high-frequency local features to enhance consistency and fidelity in long videos. Our experiments demonstrate that FreeLong significantly outperforms existing models, achieving superior temporal consistency and video fidelity. Our experiments show that FreeLong significantly outperforms existing models, achieving better temporal consistency and video fidelity. FreeLong also supports coherent multi-prompt generation, offering a practical solution for high-quality long video creation without extensive retraining.

7 Acknowledgments

This work is supported by National Science and Technology Major Project (2022ZD0117802) and the National Natural Science Foundation of China (U2336212). This work is partially supported by the Fundamental Research Funds for the Central Universities (Grant Number: 226-2024-00058).

References

[1] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[2] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

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

[4] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[5] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[6] Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. Flowzero: Zero-shot text-to-video synthesis with llm-driven dynamic scene syntax. *arXiv preprint arXiv:2311.15813*, 2023.

[7] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Eva: Zero-shot accurate attributes and multi-object video editing. *arXiv preprint arXiv:2403.16111*, 2024.

[8] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[9] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[10] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. Accessed: 2024-05-09.

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

[12] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

[13] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. *arXiv preprint arXiv:2406.04277*, 2024.

[14] 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. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.

[15] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023.

[16] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024.

[17] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024.

[18] Yu Lu, Feiyue Ni, Haofan Wang, Xiaofeng Guo, Linchao Zhu, Zongxin Yang, Ruihua Song, Lele Cheng, and Yi Yang. Show me a video: A large-scale narrated video dataset for coherent story illustration. *IEEE Transactions on Multimedia*, 2023.

 https://doi.org/10.52202/079017-4177

[19] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

[20] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[21] Yitong Li, Martin Renqiang Min, Dinghan Shen, David E. Carlson, and Lawrence Carin. Video generation from text. *CoRR*, abs/1710.00421, 2017.

[22] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David E. Carlson, and Jianfeng Gao. Storygan: A sequential conditional GAN for story visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6329–6338. Computer Vision Foundation / IEEE, 2019.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014.

[24] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *CoRR*, abs/2211.11018, 2022.

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[26] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[30] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022.

[31] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022.

[32] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.

[33] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

[34] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023.

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[36] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.

[37] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[38] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[39] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[40] Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects. *arXiv preprint arXiv:2403.16407*, 2024.

[41] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

[42] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[44] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[45] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[46] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020.

[47] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[48] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024.

 https://doi.org/10.52202/079017-4177

A Appendix

A.1 Social Impacts

It is important to consider the potential ethical implications of our approach, which is typical in generative models. By incorporating Video Diffusion Model [1, 2] into our methodology, there is a chance that our system may also inherit the biases present in these models. Additionally, we need to be aware of the potential risks, including the generation of deceptive, harmful, or discriminatory content.

A.2 Limitation

Despite its significant advancements, FreeLong has several limitations. Temporal flickering can still occur in extended sequences, affecting the visual consistency over prolonged videos. Additionally, handling dynamic scene changes where context and content vary significantly remains challenging, as the current model may struggle to adapt to rapidly changing scenarios. Nonetheless, FreeLong represents a promising approach to training-free long-form text-to-video generation, offering significant improvements in consistency and fidelity despite these challenges.

A.3 Code used and License

All used codes and their licenses are listed in Table 2.

Table 2: The used codes and license.

URL	Citation	License
`https://github.com/Vchitect/LaVie`	[1]	Apache License 2.0
`https://github.com/huggingface/diffusers`	[47]	Apache License 2.0
`https://github.com/AILab-CVC/VideoCrafter`	[2]	Apache License 2.0
`https://github.com/modelscope/modelscope`	[5]	Apache License 2.0

A.4 More Qualitative Results

We add more video generation results in Figure 9

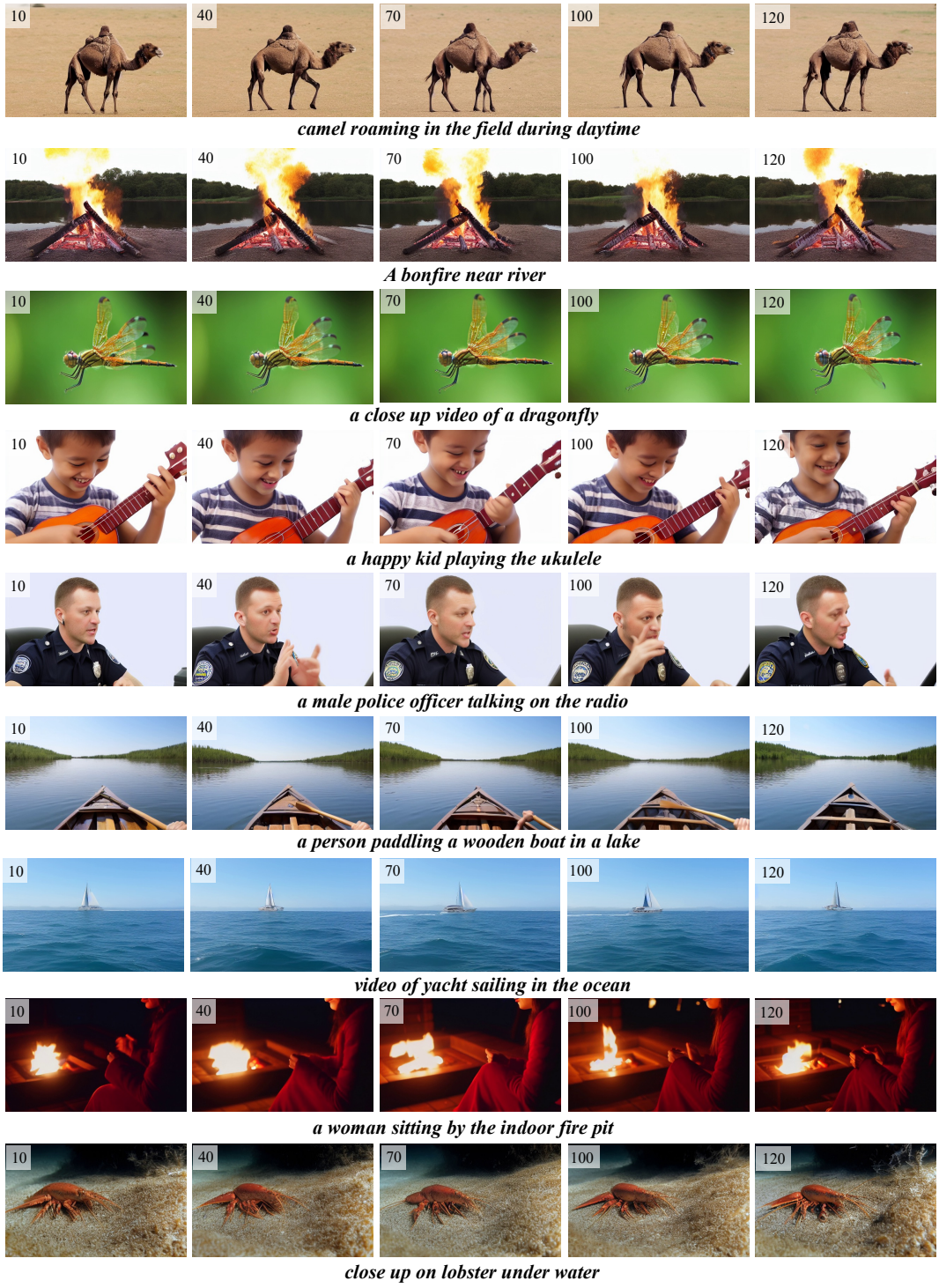


Figure 9: **More Long Video Generation Results.**

https://doi.org/10.52202/079017-4177

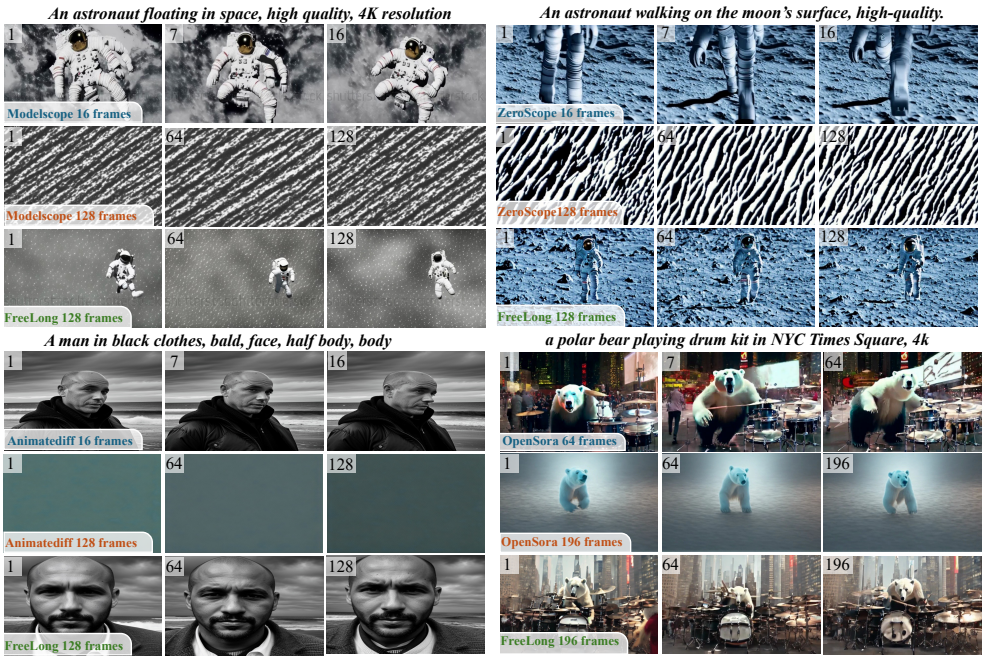


Figure 10: **Generalization to Other Base Models.** FreeLong can be easily integrated into various video diffusion frameworks by replacing the temporal attention with SpectralBlend-TA, enabling these models to generate consistent long videos with high fidelity.

A.5 Generalization to Other Base Models

Our FreeLong method is designed to be model-agnostic and can be seamlessly integrated into various pre-trained short video diffusion models. To validate this, we apply FreeLong to other state-of-the-art video diffusion frameworks beyond LaVie [1] and VideoCrafter [2]. Specifically, we incorporate FreeLong into models like ModelScope [5], Animatediff [4] and OpenSora [48], replacing their original temporal attention mechanisms with our proposed SpectralBlend Temporal Attention. As illustrated in Figure 10, FreeLong successfully enhances these models' capabilities to generate consistent long videos with high fidelity. The generated videos maintain temporal coherence across extended sequences while preserving fine-grained spatial-temporal details. The successful integration and performance improvement across different models highlight the generalizability of our approach. This flexibility allows researchers and practitioners to extend the capabilities of various short video diffusion models without additional training.

https://doi.org/10.52202/079017-4177

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We carefully described our contributions in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the appendix, we discussed our limitations, societal impact, and directions for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-4177

Justification: This paper is not about theory.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided details about the methodology and implementation in the main paper and appendix. We also upload the main code as supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released our code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We present the experimental setup and details in the main paper and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

 https://doi.org/10.52202/079017-4177

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We introduce the used computer resources in the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We carefully reviewed the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the appendix, we discussed our limitations, societal impact, and directions for future work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited related papers.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

 https://doi.org/10.52202/079017-4177

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.