FreeSpec: Training-Free Long Video Generation via Singular-Spectrum Reconstruction

Fangda Chen[1], **Shanshan Zhao**[2], **Longrong Yang**[2]
Chuanfu Xu[1], **Zhigang Luo**[3], **Long Lan**[1*]

[1]College of Computer Science and Technology,
National University of Defense Technology
[2]Alibaba International Digital Commerce
[3]Xiangjiang Laboratory, Changsha, Hunan, China

{fdchen.nudt, xuchuanfu, long.lan}@nudt.edu.cn
sshan.zhao00@gmail.com, yanglongrong.ylr@alibaba-inc.com, Zhigang_luo@sina.com

Abstract

Video diffusion models perform well in short-video synthesis, but their training-free extension to long videos often suffers from content drift, temporal inconsistency, and over-smoothed dynamics. Existing methods improve temporal consistency by combining a global branch with a local branch, but they often further decompose appearance consistency and temporal dynamics within each branch using predefined criteria. This assignment is unreliable when appearance and action progression are tightly coupled, such as in camera motion and sequential motion. We analyze the video temporal extension issue from a singular-spectrum perspective and reveal pronounced spectral concentration under enlarged self-attention windows: spectral energy becomes dominated by a few low-rank singular directions, preserving coarse structure but suppressing high-rank spatial details and motion-rich temporal variations. To mitigate this problem, we propose **FreeSpec**, a training-free spectral reconstruction framework for long-video generation. FreeSpec decomposes global and local features with singular value decomposition, and uses the global branch as low-rank spectral guidance and the local branch as a high-rank reconstruction basis. This spectrum-level fusion avoids assigning appearance and motion to mutually exclusive predefined components, preserving long-range consistency while better retaining spatial details and temporal dynamics. Experiments on Wan2.1 and LTX-Video demonstrate that FreeSpec improves long-video generation, especially for temporal dynamics, while maintaining strong visual quality and temporal consistency. Project page: `https://fdchen24.github.io/FreeSpec-Website`.

1 Introduction

Video diffusion models [1–8] have achieved remarkable progress in short-video generation, producing realistic appearances, coherent motion, and strong text alignment over dozens of frames. However, directly training long-video diffusion models remains prohibitively expensive [9–11], due to the demand for large-scale long-duration video-text data, increased memory consumption, and substantial computation for modeling extended temporal dependencies. Therefore, a practical alternative is to extend pretrained short-video models to long-range synthesis in a training-free manner.

However, training-free long-video generation remains challenging because pretrained short-video models are optimized within a limited temporal receptive field and often suffer from content drift,

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: **Long-video examples generated on Wan2.1 [7] with $4\times$ the native training length.** Existing training-free methods preserve stable appearance but may weaken continuous camera trajectories in the forest case and collapse sequential motocross actions into repetitive motion.



Figure 2: **Spectral and qualitative analysis of enlarged self-attention windows.** Here, $W = f \times h \times w$ denotes the native self-attention token length, where f, h, and w are the temporal length, height, and width of the video latent, respectively. (a) shows the effective-rank dynamics across denoising timesteps under different window sizes. (b) reports the effective rank at representative timesteps, showing lower feature rank under enlarged windows. (c) and (d) visualize low-rank reconstructions in spatial frames and temporal optical flow, respectively. The leading singular components preserve coarse structure but smooth spatial details and weaken temporal variations.

temporal inconsistency, repetitive patterns, or visual blurring when extrapolated to longer sequences. Recent studies tackle these limitations by revising the inference protocols of pretrained short-video models [12–20]. For example, Gen-L-Video [12], FreeNoise [13], and FIFO-Diffusion [14] extend the temporal horizon with overlapping segments, noise rescheduling, or queue-based denoising. RIFLEx [15], LongDiff [16], and FreeLOC [17] improve length extrapolation by modifying positional encoding or correcting context-length distribution shifts.

More closely, FreeLong [18] introduces global-local fusion in the frequency domain, while FreePCA [20] uses Principal Component Analysis (PCA) to select global or local components for stabilizing long-range generation. These methods improve long-range consistency by decomposing appearance consistency and temporal dynamics within each branch using predefined criteria. However, such decomposition can be unreliable because appearance and motion are tightly coupled in video features. For example, FreeLong [18] associates low-frequency signals in the global branch with appearance consistency and high-frequency signals in the local branch with motion details. Nevertheless, slow camera panning and long-range action evolution may also appear in low-frequency components, causing local low-frequency motion cues to be masked by global low-frequency appearance signals. As shown in Fig. 1, this limitation weakens continuous viewpoint changes in the forest case and collapses the motocross sequence into repetitive riding patterns. Overall, existing training-free methods remain limited in preserving rich temporal dynamics for long videos.

To address these challenges, we analyze the distributional properties of global- and local-branch latents. Previous studies [18–20] show that directly extending the generation length often causes

visual blurring and degraded temporal dynamics. Inspired by recent analyses of rank collapse in long-context Transformers [21, 22], we analyze this issue from a singular-spectrum perspective. As shown in Fig. 2, the effective rank increases with long denoising timesteps, but drops rapidly when the attention window is enlarged from the native token length W to $2W$, $3W$, and $4W$. This indicates that longer contexts exhibit stronger **spectral concentration**, where spectral energy is dominated by a few low-rank singular directions. We further reconstruct video features using only the low-rank components to examine this effect. The results show that these components mainly preserve coarse layout and long-range structure, while discarding fine spatial details and motion-rich temporal variations. The same spectral-concentration trend is observed on Wan2.2-5B, as reported in Appendix Sec. B.

We therefore propose FreeSpec, a training-free singular-spectrum reconstruction framework for long-video generation. FreeSpec mitigates spectral concentration by using the full-window global branch only as low-rank spectral guidance, while preserving the small-window local branch as the high-rank singular basis. Specifically, FreeSpec decomposes the global and local representations with Singular Value Decomposition (SVD) to obtain their singular values and bases. It then performs singular spectrum modulation, injecting global guidance by adjusting singular-value energy. The modulation is controlled by two factors. Timestep-aware modulation emphasizes global structural guidance in early denoising stages and reduces it in later stages for detail and motion refinement. Rank-aware modulation applies stronger global guidance to low-rank singular directions for coarse structure, while preserving higher-rank components for details and motion variations. Finally, the modulated spectrum is reconstructed under the local singular basis, preventing the output from being dominated by the low-rank, over-smoothed global basis.

We compare FreeSpec with its closest global–local fusion methods. FreeLong [18] and FreePCA [20] fuse predefined frequency and PCA components, respectively, whereas FreeSpec uses rank-dependent spectral weighting and local-basis reconstruction without explicitly separating appearance and motion. See Appendix C for details. Our contributions are summarized as follows:

- We provide a singular-spectrum analysis revealing pronounced spectral concentration under enlarged self-attention windows and demonstrating that low-rank reconstructions suppress spatial details and temporal variations.
- We propose FreeSpec, a training-free spectral reconstruction framework that injects global guidance via timestep- and rank-aware modulation and reconstructs features under local singular basis.
- We validate FreeSpec on Wan2.1 and LTX-Video under $4\times$ length extension, improving temporal smoothness and dynamic degree while maintaining competitive consistency and quality.

2 Related Work

2.1 Text-to-Video Generation Models

Text-to-video generation has advanced rapidly with diffusion-based generative models. Early methods typically adapt pre-trained text-to-image models by adding temporal modules to capture inter-frame dependencies, as shown in VideoLDM [1], LaVie [2], Stable Video Diffusion [23], AnimateDiff [24], and VideoCrafter [3, 25]. Recent progress has shifted toward diffusion transformers, where CogVideoX [4], Mochi [26], MovieGen [27], HunyuanVideo [5], LTX-Video [6], and Wan [7] demonstrate stronger scalability, motion modeling, and text-video alignment. Despite these advances, most video diffusion models are still trained on short clips and remain limited to several-second generation. Extending them to longer videos is non-trivial because long sequences introduce heavier memory costs and more complex temporal dependencies.

2.2 Long Video Generation

Long-video generation has been studied under both training-based and training-free paradigms. Training-based methods [9, 10, 28–30], such as StreamingT2V [9], Vidu [10], and autoregressive or next-clip prediction frameworks [11, 31], require costly retraining or fine-tuning. In contrast, training-free methods adapt pretrained short-video models only at inference time. Gen-L-Video [12], FreeNoise [13], and FIFO-Diffusion [14] extend generation length using overlapping segments, noise

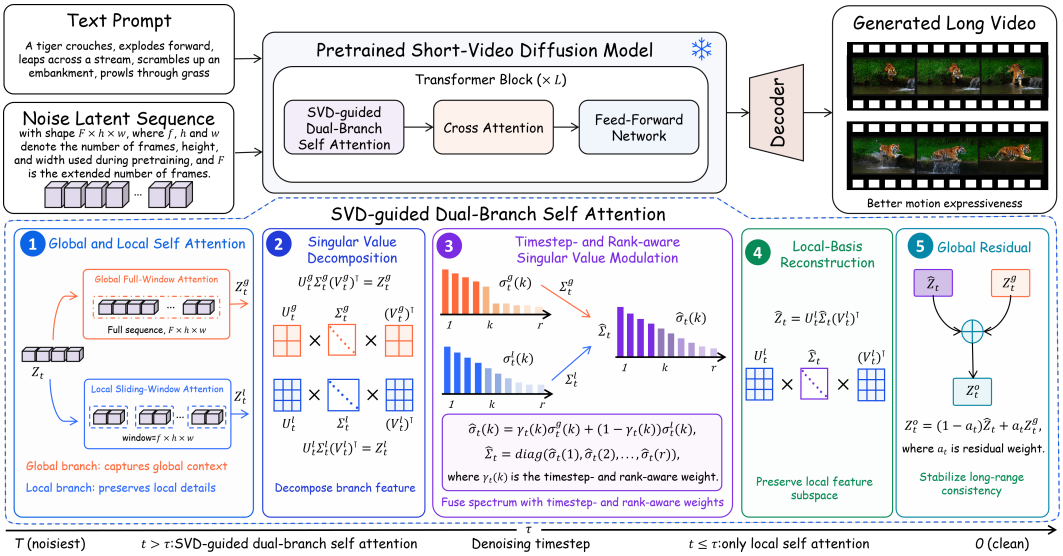

Figure 3: **Overview of FreeSpec.** FreeSpec extends a frozen short-video diffusion model to long-video generation by replacing self-attention with SVD-guided dual-branch self-attention during inference. It combines global full-window guidance and local sliding-window priors through singular-spectrum modulation, followed by local-basis reconstruction and a lightweight global residual.

rescheduling, or queue-based denoising, but their local-window design limits long-range consistency. RIFLEx [15] and LongDiff [16] improve length extrapolation by modifying positional embeddings or temporal position mappings. FreeLOC [17] further formulates long-video generation as frame-level relative-position and context-length out-of-distribution problems, and introduces layer-adaptive correction. These methods improve length extrapolation and temporal stability, yet offer limited control over the preservation of temporal dynamics. Closely related to our work, FreeLong [18] and FreeLong++ [19] adopt branch fusion in the frequency domain. They combine low-frequency global components with high-frequency local components to improve long-range stability. FreePCA [20] further replaces frequency-band blending with temporal PCA subspace blending, selecting cosine-similar components from the global branch for consistency and dissimilar components from the local branch for motion preservation. These methods demonstrate the effectiveness of global-local inference for training-free long-video generation. However, their fusion still depends on predefined decomposition criteria, which can be unreliable when structural consistency and temporal dynamics are entangled. FreeSpec fuses global and local features through rank-dependent spectral weighting, without explicitly separating appearance and motion into predefined frequency bands or PCA components. It uses the global branch as spectral guidance and reconstructs features in the local singular basis to preserve spatial details and temporal dynamics.

3 Method

3.1 Overview

FreeSpec is motivated by the spectral concentration observed in enlarged attention windows. As shown in Fig. 3, FreeSpec keeps the pretrained short-video diffusion model frozen and replaces self-attention layers with SVD-guided dual-branch self-attention during inference. For each replaced layer, FreeSpec constructs a global full-window branch and a local sliding-window branch, decomposes their outputs with SVD, injects global information through timestep- and rank-aware singular-spectrum modulation, and reconstructs the fused representation under the local singular basis. This design allows global guidance to regulate low-rank structural energy while preserving high-rank spatial-temporal variations from the local branch. Following FreePCA [20], FreeSpec applies dual-branch fusion only in early denoising stages and switches to local-only inference in later steps. This strategy reduces computation with little impact on generation quality in practice. The following sections introduce the main steps of SVD-guided dual-branch self-attention.

3.2 Dual-Branch Self-Attention

Directly applying full-window attention to long videos provides long-range context, but the resulting features exhibit pronounced spectral concentration and spatial-temporal over-smoothing. In contrast, sliding-window attention remains close to the pretrained short-video attention regime and better preserves local details and motion-rich variations, but lacks long-range consistency. We therefore adopt a dual-branch strategy that combines a full-window global branch with a sliding-window local branch, using the former for long-range structural guidance and the latter for local detail and temporal dynamics. Let F, h, w denote the temporal length, height, and width of the long-video latent, respectively, with $F \gg f$, where f is the native temporal length used during pretraining. We then define the native token window as $W = f \times h \times w$.

For query token i, local attention only attends to key-value tokens within its native neighborhood:

$$A_t^l(i,j) = \begin{cases} \mathrm{Softmax}\left(\frac{\mathbf{Q}_{t,i}\mathbf{K}_{t,j}^\top}{\sqrt{d}}\right), & |i-j| \le W, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In contrast, global attention is computed over the entire generated sequence:

$$A_t^g(i,j) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{t,i}\mathbf{K}_{t,j}^\top}{\sqrt{d}}\right). \tag{2}$$

The two branch outputs are

$$\mathbf{Z}_t^l = A_t^l \mathbf{V}_t, \qquad \mathbf{Z}_t^g = A_t^g \mathbf{V}_t. \tag{3}$$

Here, $\mathbf{Q}_t$, $\mathbf{K}_t$, and $\mathbf{V}_t$ are query, key, and value features at timestep t, and d is the feature dimension. The local branch stays close to the pretrained short-video attention regime and preserves fine details and motion-rich variations, while the global branch provides long-range context for consistency.

3.3 SVD-guided Spectral Fusion

Singular Value Decomposition. Our analysis reveals pronounced spectral concentration under enlarged attention windows, where spectral energy is dominated by a few low-rank singular directions. Low-rank reconstructions preserve coarse structure while suppressing fine spatial details and temporal variations. To explicitly operate on this spectrum, FreeSpec decomposes the global and local attention outputs with SVD:

$$\mathbf{U}_t^g \mathbf{\Sigma}_t^g (\mathbf{V}_t^g)^\top = \mathbf{Z}_t^g, \tag{4}$$

$$\mathbf{U}_t^l \mathbf{\Sigma}_t^l (\mathbf{V}_t^l)^\top = \mathbf{Z}_t^l. \tag{5}$$

Here, $\mathbf{U}_t^g$ and $\mathbf{V}_t^g$ are the singular-vector matrices of the global output $\mathbf{Z}_t^g$, and $\mathbf{U}_t^l$ and $\mathbf{V}_t^l$ are those of the local output $\mathbf{Z}_t^l$. The diagonal matrices $\mathbf{\Sigma}_t^g$ and $\mathbf{\Sigma}_t^l$ contain the corresponding singular values, which describe the spectral energy distribution of each branch. This decomposition enables singular-value modulation, allowing global structural guidance to be injected selectively.

Timestep- and Rank-aware Singular Value Modulation. Since low-rank singular directions mainly encode coarse structure while high-rank components preserve details and motion variations, and denoising models typically form global structure in early steps and refine details in later steps, global guidance should be introduced selectively rather than uniformly. FreeSpec therefore modulates singular values according to both denoising timestep and rank position.

Given a split timestep τ, for $\tau < t < T$, where T is the maximum timestep, the normalized progress within the fusion stage is

$$p_t = \frac{T-t}{T-\tau}. \tag{6}$$

Since the spectral distribution evolves nonlinearly along denoising timesteps, we use a monotonic exponential schedule to gradually shift the fusion from global structural guidance to local detail:

$$w_t^l = \frac{1-\exp(-\alpha p_t)}{1-\exp(-\alpha)}, \qquad w_t^g = 1 - w_t^l, \tag{7}$$

where α controls the transition speed. Early denoising steps use stronger global guidance for structure formation, while later steps rely more on the local branch for detail and motion refinement.

Motivated by the rapid rank reduction under enlarged attention windows, we introduce a rank-aware decay to concentrate global guidance on low-rank singular components and suppress its influence on high-rank components. Let r be the number of singular values. The global coefficient for the k-th singular value is

$$\gamma_t(k) = w_t^g \exp\left(-\beta \frac{k}{r}\right), \tag{8}$$

where β controls the decay rate along the spectrum. Let $\sigma_t^g(k)$ and $\sigma_t^l(k)$ denote the k-th singular values of the global and local branches. The fused singular value is computed as

$$\hat{\sigma}_t(k) = \gamma_t(k)\sigma_t^g(k) + (1 - \gamma_t(k))\,\sigma_t^l(k). \tag{9}$$

The fused diagonal matrix is then

$$\hat{\mathbf{\Sigma}}_t = \mathrm{diag}\left(\hat{\sigma}_t(1), \hat{\sigma}_t(2), \ldots, \hat{\sigma}_t(r)\right). \tag{10}$$

$\gamma_t(k)$ decreases as denoising timestep t decreases and rank position k increases. Therefore, the fused singular value $\hat{\sigma}_t(k)$ shifts toward the local singular value $\sigma_t^l(k)$ at later denoising steps and higher-rank positions. This design injects global structural guidance mainly into low-rank singular components and early denoising stages. For high-rank components and later timesteps, the fused spectrum is dominated by the local branch, preserving spatial details and temporal variations.

Local-Basis Reconstruction. After singular-spectrum modulation, the remaining question is which singular basis should reconstruct the fused feature. Since the global branch is more affected by spectral concentration, using its basis may project the output toward a low-rank and over-smoothed representation. FreeSpec therefore reconstructs the fused feature under the local singular basis:

$$\hat{\mathbf{Z}}_t = \mathbf{U}_t^l \hat{\mathbf{\Sigma}}_t (\mathbf{V}_t^l)^\top. \tag{11}$$

This local-basis reconstruction preserves high-rank spatial-temporal variations from the local branch, while the modulated singular values still carry controlled global structural guidance.

Global Residual. Local-basis reconstruction preserves high-rank details and motion dynamics, but the reconstructed feature may still lose part of the full-window structural signal. FreeSpec therefore adds a lightweight global residual:

$$a_t = a_0 + a_1 w_t^g, \tag{12}$$

$$\mathbf{Z}_t^o = (1 - a_t)\hat{\mathbf{Z}}_t + a_t \mathbf{Z}_t^g. \tag{13}$$

Here, a_0 and a_1 control the residual strength, and a_t is the timestep-dependent residual weight. The final output $\mathbf{Z}_t^o$ keeps the local-basis reconstruction as the main representation while retaining a small amount of global full-window information for long-range stabilization.

Similar to FreePCA [20], FreeSpec performs SVD-guided dual-branch self-attention only for $\tau < t < T$, corresponding to the early structure-forming denoising stage. For $t \leq \tau$, it uses the local branch only. This schedule improves inference efficiency.

4 Experiments

This section presents the main $4\times$-length comparisons, component ablations, and autoregressive system comparisons. The appendix provides spectral analyses, method comparisons, implementation and efficiency details, hyperparameter and schedule sensitivity, length and backbone generalization, and additional qualitative results.

4.1 Experimental Setup

Datasets. Following recent training-free long video generation methods [18–20], we evaluate our method on long text-to-video generation using 100 enhanced prompts from VBench-Long [32].

Compared Methods. We compare FreeSpec with representative training-free long video generation baselines. **Direct** directly applies full self-attention to the extended video sequence without any adaptation. **Sliding Window** restricts self-attention to a local window whose length is the same as the training length W. **FreeNoise** [13] extends short-video diffusion models through noise rescheduling

Table 1: **Quantitative comparison on Wan2.1-1.3B ($4\times$).** The best result is highlighted in **bold**, and the second-best result is underlined. Inference time denotes the runtime on an A100 GPU.

Method	Subj.-Con.	Bkgd.-Con.	Mot.-Smth.	Dyn.-Deg.	Aes.-Qual.	Img.-Qual.	Infer. Time
Direct	**0.9865**	**0.9752**	0.9876	0.0468	0.5858	0.6274	∼38 mins
Sliding Window	0.9791	0.9677	0.9887	**0.3955**	0.6018	0.6424	∼25 mins
FreeNoise	0.9777	0.9667	0.9847	0.2395	0.6074	**0.6612**	∼25 mins
FreeLong	0.9849	0.9731	0.9856	0.2514	0.6057	0.6374	∼61 mins
FreePCA	0.9857	0.9734	0.9853	0.2177	0.6005	0.6372	∼55 mins
FreeLOC	0.9783	0.9717	0.9846	0.2405	0.5964	0.6432	∼63 mins
FreeSpec (Ours)	0.9863	0.9745	**0.9902**	0.3200	**0.6104**	0.6415	∼56 mins

Table 2: **Quantitative comparison on LTX-Video-2B-dev ($4\times$).** The best result is highlighted in **bold**, and the second-best result is underlined. Inference time denotes the runtime on an A100 GPU.

Method	Subj.-Con.	Bkgd.-Con.	Mot.-Smth.	Dyn.-Deg.	Aes.-Qual.	Img.-Qual.	Infer. Time
Direct	**0.9726**	**0.9682**	0.9800	0.1933	0.5448	0.6468	∼13 mins
Sliding Window	0.9559	0.9620	0.9795	0.2311	0.5442	0.6282	∼12 mins
FreeNoise	0.9618	0.9614	0.9822	0.0556	0.5230	**0.6528**	∼12 mins
FreeLong	0.9699	0.9640	0.9804	0.0311	0.5430	0.6478	∼22 mins
FreePCA	0.9610	0.9637	0.9796	0.2111	0.5453	0.6332	∼21 mins
FreeLOC	0.9719	0.9633	0.9852	0.1500	0.5454	0.6257	∼22 mins
FreeSpec (Ours)	0.9622	0.9677	**0.9892**	**0.2478**	**0.5464**	0.6382	∼22 mins

and window-based attention fusion. **FreeLong** [18] adopts a global-local inference paradigm and fuses the low-frequency components of the global branch with the high-frequency components of the local branch in the frequency domain. **FreePCA** [20] decomposes global and local features into a shared principal subspace and progressively injects global consistency components into local features according to component-level similarity. **FreeLOC** [17] applies layer-adaptive corrections for relative-position and context-length distribution shifts.

Evaluation Metrics. We adopt VBench-based metrics [32, 33] to evaluate video consistency, temporal quality, and perceptual quality. For consistency, we use **Subject Consistency**, which measures cross-frame object appearance similarity using DINO features [34], and **Background Consistency**, which measures scene-level background stability using CLIP features [35]. For temporal quality, we report **Motion Smoothness**, which uses motion priors from a frame interpolation model to assess whether generated motion evolves smoothly [36], and **Dynamic Degree**, which quantifies motion intensity via optical flow magnitude [37]. For perceptual quality, we report **Aesthetic Quality**, which measures artistic merit using the LAION aesthetic predictor [38], and **Imaging Quality**, which evaluates frame-level visual quality using an image quality predictor [39].

4.2 Main Results

We evaluate FreeSpec on two public text-to-video backbones, Wan2.1-1.3B [7] and LTX-Video-2B-dev [6]. For each backbone, we generate videos with $4\times$ the native training length, i.e., 321 frames for Wan2.1-1.3B and 481 frames for LTX-Video-2B-dev.

Quantitative comparison. Tables 1 and 2 report the quantitative results. Direct full attention achieves strong subject and background consistency but yields low dynamic degree, consistent with over-smoothed temporal evolution under enlarged attention windows. Sliding Window preserves motion within the native temporal range, but weakens long-range consistency. FreeNoise improves visual quality on some metrics, yet its consistency and motion scores remain limited, indicating that noise rescheduling alone is insufficient to preserve spatial consistency and temporal dynamics. FreeLong and FreePCA obtain competitive consistency by using global-local fusion, but their frequency- or PCA-based decomposition can still suppress useful motion cues, leading to weaker dynamic degree. FreeLOC [17] obtains competitive consistency with layer-adaptive correction, but its Dynamic Degree remains lower than FreeSpec on both Wan2.1 and LTX-Video. In contrast, FreeSpec achieves the best motion smoothness on both backbones and improves dynamic degree over most global-local baselines, while maintaining competitive consistency and perceptual quality.

Qualitative comparison. Fig. 4 shows consistent qualitative trends. Direct full attention produces stable but nearly static videos with visual blurring, while Sliding Window introduces more motion at the cost of appearance and layout stability. FreeNoise often improves frame-level quality, but



Figure 4: **Qualitative comparison under 4× length extension on Wan2.1 and LTX-Video.**

Table 3: **Ablation results on the first 50 prompts of the evaluation set using Wan2.1-1.3B. Bold** and underline indicate the best and second-best results, respectively.

Component	Variant	Subj.-Con.	Bkgd.-Con.	Mot.-Smth.	Dyn.-Deg.	Aes.-Qual.	Img.-Qual.
Global basis	GB	**0.9931**	**0.9817**	**0.9904**	0.2182	0.6259	**0.6749**
Local basis	LB	0.9798	0.9678	0.9887	**0.4236**	0.6233	0.6611
Fixed modulation	LB + fixed modulation	0.9821	0.9696	0.9895	0.3782	0.6175	0.6471
Weighted modulation	LB + RA	0.9820	0.9696	0.9897	0.3455	0.6078	0.6212
	LB + TA	0.9810	0.9687	0.9895	0.3800	0.6196	0.6458
	LB + RA + TA	0.9809	0.9688	0.9894	0.3836	0.6189	0.6494
Lightweight residual	LB + T-GR only	0.9877	0.9757	0.9903	0.3218	**0.6287**	0.6546
	LB + RA + TA + F-GR	0.9857	0.9737	0.9898	0.3891	0.6191	0.6432
	LB + RA + TA + T-GR	0.9860	0.9742	0.9899	0.3891	0.6207	0.6462

action progression remains weak. FreeLong and FreePCA better stabilize global structure, yet their decomposition-based fusion may still produce conservative motion or incomplete temporal evolution. FreeLOC also maintains good consistency but exhibits limited motion dynamics. By contrast, FreeSpec preserves subject appearance and long-range layout while producing richer temporal changes, such as smoother human-dog interactions on Wan2.1-1.3B, and maintaining stronger long-range consistency in the relatively static landscape sequence on LTX-Video-2B-dev. These results further demonstrate that FreeSpec mitigates over-smoothing while preserving motion-rich temporal dynamics in dynamic scenes and long-range consistency in relatively static scenes.

Efficiency Analysis. FreeSpec adds two attention branches and post-attention spectral reconstruction to the frozen backbone. For $L \gg C$, where L is the token count and C is the feature dimension, each economy-size SVD has leading cost $O(LC^2)$, so native inference is dominated by dual-branch attention and SVD. On both Wan2.1-1.3B and LTX-Video-2B-dev at $4\times$ length, FreeSpec remains in the same runtime range as FreeLong, FreePCA and FreeLOC. FreeSpec's post-attention design is compatible with accelerated bidirectional systems such as LightX2V [40], enabling generation of $4\times$-length videos in under three minutes, with an inference time comparable to some autoregressive systems [11, 30]. Detailed native runtime and memory profiles are provided in Appendix Sec. E.

4.3 Ablation Study

We conduct ablations on the first 50 evaluation prompts. In table order, **GB** and **LB** denote global- and local-basis reconstruction; **LB + fixed modulation** uses equal global/local singular-value weights; **RA**, **TA**, and **RA + TA** denote rank-aware, timestep-aware, and combined adaptive modulation; **LB + T-GR only** applies the timestep-aware global residual directly to LB; and **F-GR/T-GR** denote fixed/timestep-aware residuals added after RA+TA.

Quantitative ablation. Table 3 reports the quantitative ablation results. GB achieves the strongest consistency and imaging quality but the lowest Dynamic Degree. LB increases Dynamic Degree from



Figure 5: Qualitative ablation results of FreeSpec on Wan2.1. Yellow boxes highlight representative regions with subject inconsistency, background drift, local artifacts, or weakened temporal evolution.

0.2182 to 0.4236 at the expense of consistency, while fixed modulation partially restores consistency but reduces motion and quality. RA and TA improve this trade-off along different dimensions, and their combination further improves motion and perceptual quality. Among the lightweight residual variants, LB + T-GR improves consistency but suppresses motion relative to LB. Applying F-GR or T-GR after RA+TA instead preserves the recovered dynamics while improving consistency, with T-GR achieving higher consistency than F-GR at the same Dynamic Degree. These results clarify the roles of the three components: local-basis reconstruction preserves motion dynamics; rank- and timestep-aware spectral modulation balances global guidance with motion and perceptual quality; and lightweight global residual guidance further strengthens consistency while retaining the dynamics recovered by spectral modulation.

Qualitative ablation. Fig. 5 shows the same consistency–motion trade-off qualitatively. GB maintains a stable subject appearance and background layout, but exhibits limited wing motion. LB produces clearer pose transitions and richer wing dynamics, while introducing subject inconsistency, background drift, and local artifacts. Fixed modulation, RA, TA, and RA+TA refine this trade-off in different ways, with RA+TA retaining clearer motion while improving scene stability. LB + T-GR only stabilizes the subject and background but weakens temporal evolution, whereas adding F-GR or T-GR after RA+TA preserves the recovered motion and further suppresses large-scale inconsistencies. T-GR provides the cleanest overall result among the residual variants. These observations support local-basis reconstruction with controlled global guidance and lightweight residual correction.

4.4 Autoregressive and Accelerated-System Comparisons

To assess FreeSpec's quality–efficiency trade-off against autoregressive systems and its compatibility with accelerated inference, we compare both native and LightX2V-accelerated FreeSpec with representative AR methods. All systems are generated and evaluated using the same 100 prompts, 321-frame outputs ($4\times$ length) with 1.3B-parameter generators. AR methods use official checkpoints and native sampling pipelines. Native FreeSpec uses the frozen bidirectional backbone, while its accelerated variant integrates FreeSpec into the four-step distilled LightX2V system [40].

Table 4 compares AR systems, native FreeSpec, and its LightX2V integration. CausVid achieves strong subject and background consistency, while Rolling Forcing favors motion smoothness and

Table 4: System comparison on 100 prompts. The last two rows provide a controlled FreeSpec on/off comparison within LightX2V.

System	Subj.-Con.	Bkgd.-Con.	Mot.-Smth.	Dyn.-Deg.	Aes.-Qual.	Img.-Qual.
CausVid [41]	0.9863	0.9732	0.9871	0.2809	0.6151	0.7039
Self-Forcing [11]	0.9799	0.9679	0.9904	0.2782	0.6190	<u>0.7183</u>
Rolling Forcing [30]	0.9845	0.9690	<u>0.9907</u>	0.2555	**0.6257**	**0.7360**
Infinity-RoPE [42]	0.9844	0.9708	0.9906	0.3627	0.6240	0.7154
Causal Forcing [43]	0.9684	0.9600	0.9802	**0.4909**	0.5961	0.7017
Native FreeSpec	0.9863	<u>0.9745</u>	0.9902	0.3200	0.6104	0.6415
LightX2V (FreeSpec off)	**0.9925**	**0.9787**	**0.9926**	0.3000	<u>0.6252</u>	0.7191
FreeSpec + LightX2V	<u>0.9899</u>	0.9744	0.9906	<u>0.4418</u>	0.6209	0.7077

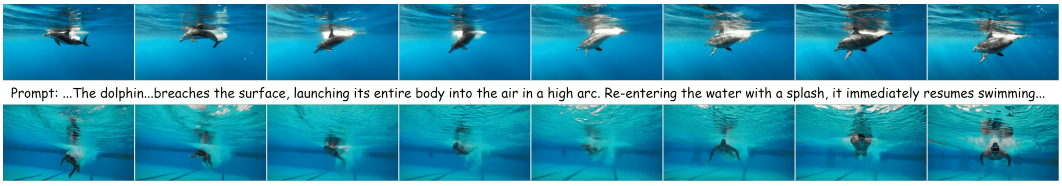

Figure 6: **Failure cases.** FreeSpec preserves temporal dynamics but fails to infer implicit scene transitions, such as underwater-to-air and diving-platform-to-pool transitions.

perceptual quality but exhibits limited dynamics. Causal Forcing achieves the highest Dynamic Degree, with lower consistency and motion smoothness; Infinity-RoPE offers a more balanced motion–quality trade-off. Native FreeSpec matches the highest AR Subject Consistency and exceeds all listed AR Background Consistency scores, although its Imaging Quality is lower. LightX2V alone achieves the strongest consistency and motion smoothness, but its Dynamic Degree remains limited. Integrating FreeSpec into LightX2V increases Dynamic Degree by 47.3% with modest reductions in the other metrics, yielding the second-highest dynamics among the evaluated systems while retaining strong consistency. These results demonstrate that FreeSpec complements distilled acceleration by improving temporal dynamics without additional FreeSpec-specific training.

5 Failure Cases and Limitations

Limitations. Although FreeSpec improves long-range temporal dynamics, it inherits the semantic limitations of the pretrained base model. In Fig. 6, the generated dolphin remains underwater instead of transitioning to air, while the diving sequence omits the platform stage. These failures highlight the lack of explicit event-level planning and reliable control over implicit scene transitions.

Broader impacts. FreeSpec may support research prototyping by extending existing video models without retraining. However, improved long-video synthesis may also facilitate deceptive media and disinformation, motivating provenance tracking, watermarking, and content moderation.

6 Conclusion

In this paper, we presented FreeSpec, a training-free spectral reconstruction framework for long-video generation. Our analysis revealed pronounced spectral concentration under enlarged self-attention windows, with feature energy dominated by a few low-rank singular directions. Low-rank reconstructions preserve coarse structure but suppress high-rank details and temporal variations, resulting in spatial smoothing and weakened dynamics. To address this, FreeSpec uses the global branch only as low-rank spectral guidance and reconstructs features under the local singular basis. Timestep- and rank-aware modulation further controls global guidance across denoising stages and singular directions. Experiments on Wan2.1 and LTX-Video demonstrate that FreeSpec achieves a better balance between long-range consistency, motion dynamics, and visual quality. However, as a training-free method, it is still limited by the semantic reasoning ability of the frozen base model, especially for implicit scene transitions in sequential actions.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.62622617, 62376282) and the Science and Technology Innovation Program of Hunan Province (No.2025RC3117).

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[2] 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. *International Journal of Computer Vision*, 133(5):3059–3078, 2025.

[3] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7310–7320, 2024.

[4] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

[5] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[6] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[7] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[8] Fangda Chen, Shanshan Zhao, Chuanfu Xu, and Long Lan. Jointtuner: Appearance-motion adaptive joint training for customized video generation. *arXiv preprint arXiv:2503.23951*, 2025.

[9] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025.

[10] 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.

[11] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[12] 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.

[13] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868, 2024.

[15] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. In *Forty-second International Conference on Machine Learning*, 2025.

[16] Zhuoling Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Longdiff: Training-free long video generation in one go. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17789–17798, 2025.

[17] Jiahao Tian, Chenxi Song, Wei Cheng, and Chi Zhang. Free-lunch long video generation via layer-adaptive ood correction. *arXiv preprint arXiv:2603.25209*, 2026.

[18] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37: 131434–131455, 2024.

[19] Yu Lu and Yi Yang. Freelong++: Training-free long video generation via multi-band spectralfusion. *arXiv preprint arXiv:2507.00162*, 2025.

[20] Jiangtong Tan, Hu Yu, Jie Huang, Jie Xiao, and Feng Zhao. Freepca: Integrating consistency information across long-short frames in training-free long video generation via principal component analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27979–27988, 2025.

[21] Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In *Forty-second International Conference on Machine Learning*, 2025.

[22] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. *arXiv preprint arXiv:2510.05554*, 2025.

[23] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[24] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[25] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[26] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

[27] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[28] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

[29] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[30] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

[31] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

[32] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

[33] 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. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[34] 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.

[35] 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.

[36] 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.

[37] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[38] LAION-AI. aesthetic-predictor, 2022. https://github.com/LAION-AI/aesthetic-predictor.

[39] 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.

[40] LightX2V Contributors. Lightx2v: Light video generation inference framework. `https://github.com/ModelTC/lightx2v`, 2025.

[41] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 22963–22974. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02138. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html`.

[42] Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pages 40256–40265. Computer Vision Foundation, 2026. URL `https://openaccess.thecvf.com/content/CVPR2026/html/Yesiltepe_Infinity-RoPE_Action-Controllable_Infinite_Video_Generation_Emerges_From_Autoregressive_Self-Rollout_CVPR_2026_paper.html`.

[43] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *CoRR*, abs/2602.02214, 2026. doi: 10.48550/ARXIV.2602.02214. URL `https://doi.org/10.48550/arXiv.2602.02214`.

A Supplementary Overview

This appendix provides spectral analyses, method comparisons, implementation and efficiency details, additional ablations, length and backbone generalization results, and qualitative comparisons.

- Sec. B analyzes spectral concentration and spectral components, including validation on Wan2.2-TI2V-5B.
- Sec. C contrasts FreeSpec with training-free global-local fusion methods, highlighting explicit appearance–motion partitioning.
- Sec. D summarizes default hyperparameters and denoising settings.
- Sec. E profiles runtime, memory, and computational bottlenecks across configurations.
- Sec. F evaluates hyperparameter and modulation-schedule sensitivity.
- Sec. G evaluates length and backbone generalization without retuning.
- Sec. H presents qualitative comparisons on complex motion and VBench-Long prompts.

B Observation and Analysis

This section analyzes how enlarged self-attention windows affect the singular spectrum of intermediate video diffusion features. We examine effective-rank evolution during denoising and use spatial-frame and optical-flow reconstructions to interpret the roles of different spectral components. We further test whether the window-dependent spectral trend persists on the larger Wan2.2-TI2V-5B backbone.

Motivated by recent spectral analyses of long-context transformers [21, 22], we examine how context extension affects intermediate video representations in diffusion-based video generation. Prior studies suggest that enlarging the attention context may lead to more uniform attention scores, excessive token aggregation, and rank collapse along the context-length dimension. Since training-free long-video generation often extends pretrained short-video models by enlarging attention windows during inference, we investigate whether similar spectral degeneration also appears in video diffusion features.

Specifically, we analyze the singular spectrum of intermediate representations under different temporal self-attention windows, including W, $2W$, $3W$, and $4W$. Here, $W = f \times h \times w$ denotes the native self-attention token length, where f, h, and w denote the temporal length, height, and width used during pretraining, respectively. Enlarging the window from W to $4W$ therefore extends the temporal receptive field from the native pretrained context to a four-times longer token context. We compute the effective rank over denoising timesteps on 20 prompts and report the mean curve with standard error. To further interpret the role of different spectral components, we reconstruct video representations using different rank ratios and visualize both spatial frames and temporal optical flow.

Effective rank follows temporal growth but decays with window enlargement. We first study how denoising progress and temporal inference window jointly affect the rank structure of video representations. Given singular values $\{\sigma_i\}_{i=1}^{n}$, we normalize them as $p_i = \sigma_i / \sum_j \sigma_j$, and compute the effective rank as $r_{\text{eff}} = \exp(-\sum_i p_i \log p_i)$. A larger effective rank indicates that representation energy is distributed across more singular directions, whereas a smaller value indicates stronger concentration on a few dominant components.

As shown in Fig. 2(a), for each fixed self-attention window, the effective rank increases rapidly during early denoising and then gradually approaches a plateau, forming a growth pattern over denoising timesteps. This suggests that video representations progressively activate more singular directions as generation evolves from coarse structure formation to detailed content refinement. However, this temporal growth is strongly modulated by the attention window size. Enlarging the window from W to $2W$, $3W$, and $4W$ consistently lowers the effective-rank curve, showing that longer temporal contexts exhibit stronger spectral concentration, with feature energy dominated by fewer singular components.

Fig. 2(b) further shows this window-dependent effect at representative denoising stages, including the rising, near-plateau, and post-plateau phases. Across all stages, the effective rank drops sharply when the window is enlarged from W to $2W$, and then decreases more slowly from $2W$ to $4W$, exhibiting

an exponential-like decay with respect to window size. This consistent rank decay shows that extending the temporal context concentrates feature energy into fewer dominant singular components.

Low-rank components preserve coarse structure but suppress details and motion. We then analyze the semantic role of different spectral components by reconstructing video representations with different rank ratios. As shown in Fig. 2(c), the full-rank reconstruction preserves clear spatial appearance, including object boundaries, local textures, and background details. When only the top-50% singular components are retained, the global layout and main subject structure remain recognizable, but local details become visibly smoother. This degradation becomes more pronounced in the top-20% and top-10% reconstructions, where frames preserve only coarse appearance and lose fine-grained texture.

A similar tendency can be observed in the temporal optical-flow visualization in Fig. 2(d). Full-rank reconstruction produces stronger and more structured flow responses around moving regions, while low-rank reconstructions yield weaker and less detailed flow maps, indicating that motion-rich temporal variations are largely suppressed. These observations suggest that low-rank singular components mainly encode coarse layout and long-range structure, whereas higher-rank components are important for local texture, sharp boundaries, and expressive motion dynamics.

Overall, the analysis reveals a dual effect of temporal context extension. Along denoising timesteps, video representations naturally evolve toward richer rank structures, while enlarged self-attention windows yield lower effective rank and stronger spectral concentration. While this low-rank bias may benefit long-range structural consistency, excessive concentration suppresses high-rank components essential for visual fidelity and motion richness. This motivates our spectral reconstruction strategy: instead of directly replacing local representations with global-window outputs, we inject global structural guidance through controlled singular-spectrum modulation while preserving local high-rank components for fine spatial details and dynamic temporal variations.

Spectral concentration persists on a larger backbone. To test whether the observed trend extends beyond Wan2.1-1.3B, we repeat the effective-rank diagnostic on Wan2.2-TI2V-5B using 20 prompts. Fig. A1 compares self-attention windows from W to $4W$ at the rising ($t = 963$), near-plateau ($t = 906$), and post-plateau ($t = 701$) denoising stages. Each point reports the mean effective rank, with error bars indicating the standard error of the mean.

Across all three stages, enlarged self-attention windows yield lower effective rank, with reductions of 28.2%, 30.0%, and 33.1% from W to $4W$, respectively. The largest reduction occurs from W to $2W$, followed by smaller decreases toward $4W$, reproducing the steep initial decline observed on Wan2.1. The same window-size ordering holds at all 50 denoising timesteps. These results show that window-dependent spectral concentration persists on a larger backbone, further motivating controlled spectral reconstruction.

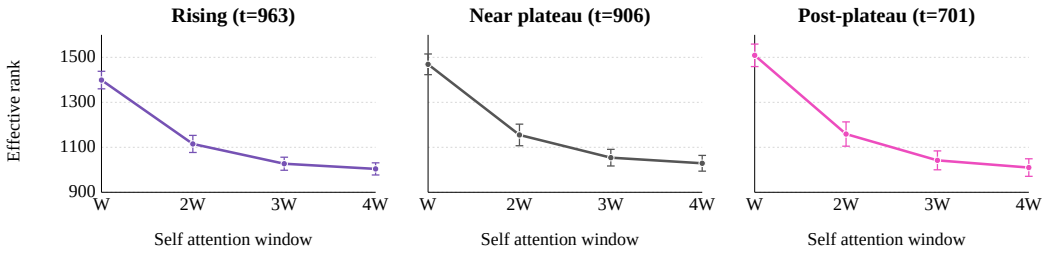

Figure A1: Effective rank under enlarged self-attention windows on Wan2.2-TI2V-5B at three denoising stages.

C Method Comparison

Fig. A2 illustrates the differences between FreeSpec and representative training-free global-local fusion methods, highlighting whether branch features are explicitly partitioned into appearance and motion components.

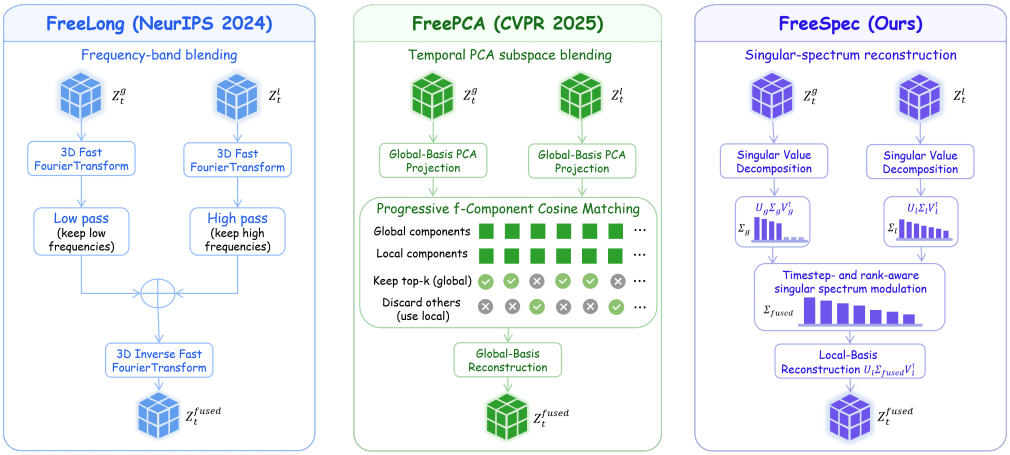

Figure A2: Visual comparison of training-free global-local fusion paradigms. FreeLong uses frequency-band fusion, FreePCA uses PCA-component selection, and FreeSpec performs singular-spectrum modulation followed by local-basis reconstruction.

D Implementation Details

This section summarizes the default implementation settings of FreeSpec. Unless otherwise specified, the main hyperparameters are fixed across backbones and evaluation cases: the timestep threshold is $\tau = 0.9$ with $T = 1.0$, the timestep transition coefficient is $\alpha = 5.0$, the rank-aware decay factor is $\beta = 5.0$, and the timestep-aware global residual schedule uses $a_0 = 0.15$ and $a_1 = 0.2$. Under the shifted 50-step Wan denoising schedule (shift 8), this threshold enables dual-branch fusion for the first 24 steps, followed by local-only attention. All five hyperparameters control inference and require no model training.

E Runtime, Memory, and Scaling

This section profiles the runtime and memory usage of FreeSpec across video lengths and backbone sizes, identifying the main computational bottlenecks and the limits of cross-configuration scaling comparisons. Table A1 characterizes the runtime and memory costs of three native FreeSpec configurations. Each result averages three profiling runs. End-to-end time includes preprocessing, text encoding, denoising, VAE decoding, and cleanup, excluding model loading and video encoding. Component percentages are relative to denoising time, and peak memory is measured from allocator statistics in GiB. The Wan2.1-1.3B $4\times$ configuration generates 321 frames at 832×480 using 50 denoising steps, seed 42, and an A100-40GB. The $8\times$ and Wan2.2-5B configurations use a 48GB GPU because the attempted runs exceeded the A100-40GB memory capacity.

Attention and SVD dominate FreeSpec's denoising cost, while spectral modulation and reconstruction add little overhead. For Wan2.1-1.3B at $4\times$ length, SVD is the largest individual component (40.1%), and the two attention branches together account for 48.0%. At $8\times$ length, attention accounts for 71.0%, including 46.6% from global attention, whereas SVD accounts for 15.1%. For Wan2.2-5B, SVD remains a major cost at 41.8%. Singular modulation, local reconstruction, and the global residual each account for at most 0.4% across all three configurations. These profiles identify attention and decomposition as the main targets for acceleration, although hardware differences prevent attributing runtime or memory changes solely to video length or model size.

F Additional Experiments

This section provides additional experimental results beyond the main paper. We first analyze the sensitivity of FreeSpec to key hyperparameters, then compare modulation schedules to assess their effects on temporal consistency, motion dynamics, and visual quality.

Table A1: FreeSpec runtime and peak memory. Component times are in seconds, with percentages of denoising time in parentheses. Length multipliers indicate extension beyond the native temporal window.

Metric / component	Wan2.1-1.3B $4\times$ length A100-40GB	Wan2.1-1.3B $8\times$ length 48GB GPU	Wan2.2-5B $4\times$ length 48GB GPU
Global attention	780.60 (22.5%)	3341.84 (46.6%)	885.87 (16.1%)
Local attention	884.99 (25.5%)	1750.75 (24.4%)	1139.15 (20.7%)
SVD	1391.31 (40.1%)	1081.89 (15.1%)	2292.30 (41.8%)
Singular modulation	4.05 (0.1%)	0.35 ($<0.1\%$)	0.29 ($<0.1\%$)
Local reconstruction	10.42 (0.3%)	25.39 (0.4%)	19.39 (0.4%)
Global residual	4.91 (0.1%)	13.27 (0.2%)	8.44 (0.2%)
Other denoising	396.59 (11.4%)	960.39 (13.4%)	1144.93 (20.9%)
End-to-end time (min)	58.41	120.82	94.15
Peak allocated memory (GiB)	25.39	21.38	41.40
Peak reserved memory (GiB)	30.95	33.92	46.79

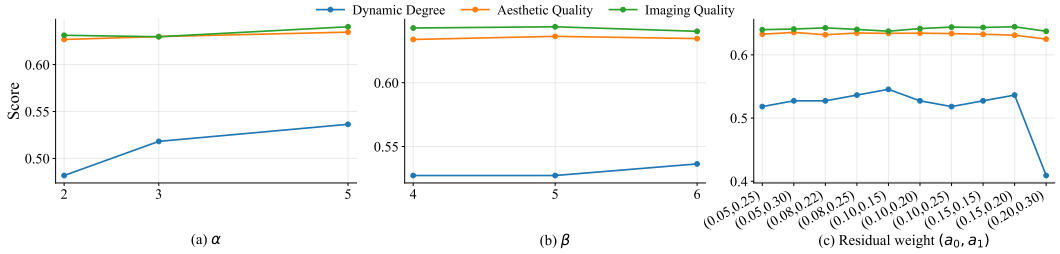

Figure A3: Hyperparameter sensitivity of FreeSpec. We vary α, β, and $a_t = a_0 + a_1 w_t^g$, and report dynamic degree, aesthetic quality, and imaging quality.

F.1 Hyperparameter Analysis

We analyze the sensitivity of FreeSpec to three key hyperparameters: the timestep transition coefficient α, the rank-decay coefficient β, and the residual weight $a_t = a_0 + a_1 w_t^g$. As shown in Fig. A3, we focus on Dynamic Degree, Aesthetic Quality, and Imaging Quality, which reflect motion richness and perceptual quality.

For α, increasing it from 2 to 5 consistently improves dynamic degree, with mild gains in aesthetic and imaging quality. This suggests that a sharper transition from global guidance to local dynamics better preserves motion richness in later denoising stages. Although $\alpha = 5$ slightly reduces consistency compared with $\alpha = 2$ and $\alpha = 3$, it achieves the best overall trade-off among motion dynamics and perceptual quality, and is therefore used as the default setting.

For β, performance remains stable within the tested range. Increasing β from 4 to 6 slightly improves dynamic degree, while aesthetic and imaging quality change only marginally. This indicates limited sensitivity to the rank-decay coefficient within the tested range, and moderate rank-aware modulation is sufficient to suppress global influence on high-rank components.

For the residual weight, the trade-off is more evident. Moderate settings, such as $(0.08, 0.25)$, $(0.10, 0.15)$, and $(0.15, 0.20)$, maintain competitive perceptual quality and motion dynamics. In contrast, an overly strong residual, such as $(0.20, 0.30)$, sharply reduces dynamic degree and also degrades aesthetic and imaging quality. This confirms that excessive global residual guidance can overwrite the local branch and weaken temporal dynamics.

F.2 Modulation Schedules

To assess how the modulation schedule affects the balance between temporal consistency, motion dynamics, and visual quality, we compare linear, cosine, and exponential schedules on the same first 50 Wan2.1-1.3B prompts. As shown in Table A2, exponential modulation achieves the highest

Table A2: **Modulation-schedule ablation on the first 50 prompts of the evaluation set using Wan2.1-1.3B. Bold** and underline indicate the best and second-best results, respectively.

Schedule	Subj.-Con.	Bkgd.-Con.	Mot.-Smth.	Dyn.-Deg.	Aes.-Qual.	Img.-Qual.
Linear	0.9874	<u>0.9757</u>	**0.9905**	0.2964	<u>0.6214</u>	<u>0.6390</u>
Cosine	**0.9877**	**0.9761**	<u>0.9905</u>	<u>0.3073</u>	**0.6215**	0.6366
Exponential	<u>0.9860</u>	0.9742	0.9899	**0.3891**	0.6207	**0.6462**

Dynamic Degree (0.3891, versus 0.2964 for linear and 0.3073 for cosine) and Imaging Quality, with only small reductions in consistency, motion smoothness, and Aesthetic Quality. These results support our choice of exponential modulation for its favorable balance between stronger dynamics and preserved consistency and quality.

G Additional Length and Backbone Generalization

This section evaluates whether FreeSpec preserves motion under longer temporal extensions and transfers to a larger backbone without hyperparameter retuning. We extend Wan2.1-T2V-1.3B to $8\times$ its native temporal window and evaluate Wan2.2-TI2V-5B at $4\times$, testing generalization beyond the main experimental settings.

Both experiments use a fixed 20-prompt subset and the default FreeSpec hyperparameters. The subset comprises the first 20 entries of the ordered 100-prompt VBench-Long list, fixed before inspecting any generated videos or metrics, with no output-based selection or replacement. Wan2.1 generates 641 frames at 832×480 and 16 FPS, while Wan2.2 generates 481 frames at 1280×704 and 24 FPS. We retain $(\alpha, \beta, a_0, a_1, \tau, T) = (5, 5, 0.15, 0.20, 0.9, 1.0)$ without retuning.

Comparisons are controlled within each backbone, with only the attention and fusion strategy varying across methods. Direct, FreeLong, FreePCA, and FreeSpec share the same prompt indices, checkpoint, seed 42, Gaussian-noise initialization, and VBench-Long evaluation protocol. Wan2.1 uses UniPC with 50 steps, shift 8, guidance scale 6, and prompt extension disabled. Wan2.2 uses a shared generation pipeline with dense global attention, split timestep 900, and sparse mode disabled. The native local latent windows are 21 and 31 for Wan2.1 and Wan2.2, respectively.

Table A3: Length and backbone generalization without retuning on a fixed 20-prompt subset. Length multipliers are relative to each backbone's native temporal window. Best and second-best scores within each setting are bold and underlined.

Method	Subj.-Con.	Bkgd.-Con.	Mot.-Smth.	Dyn.-Deg.	Aes.-Qual.	Img.-Qual.
Wan2.1-T2V-1.3B: $8\times$ length extension 641 frames, 832×480, 16 FPS						
Direct	**0.9890**	**0.9764**	<u>0.9887</u>	0.0524	0.5995	**0.6460**
FreeLong	0.9850	0.9711	0.9857	<u>0.1833</u>	0.6038	0.6135
FreePCA	<u>0.9873</u>	<u>0.9741</u>	0.9862	0.1286	<u>0.6058</u>	<u>0.6216</u>
FreeSpec	0.9868	0.9740	**0.9906**	**0.2381**	**0.6064**	0.6194
Wan2.2-TI2V-5B: $4\times$ length extension 481 frames, 1280×704, 24 FPS						
Direct	**0.9854**	**0.9816**	0.9888	0.1545	0.6136	0.5702
FreeLong	<u>0.9841</u>	<u>0.9795</u>	0.9888	<u>0.1864</u>	0.6077	**0.5889**
FreePCA	<u>0.9806</u>	0.9772	<u>0.9894</u>	0.1636	<u>0.6169</u>	<u>0.5799</u>
FreeSpec	<u>0.9806</u>	0.9786	**0.9945**	**0.2045**	**0.6187**	0.5592

FreeSpec achieves the highest Motion Smoothness, Dynamic Degree, and Aesthetic Quality in both settings, demonstrating that its motion benefits persist without retuning. As shown in Table A3, Direct retains the strongest subject and background consistency but produces the lowest Dynamic Degree, particularly at $8\times$ length. FreeLong and FreePCA improve dynamics over Direct, yet remain below FreeSpec: at $8\times$, FreeSpec reaches 0.2381 versus 0.1833 and 0.1286, respectively; on Wan2.2, it reaches 0.2045 versus 0.1864 and 0.1636. These gains come with modestly lower consistency than Direct and, on Wan2.2, lower Imaging Quality than all three baselines. The results therefore

support motion preservation across longer outputs and a larger backbone, while revealing a remaining imaging-quality trade-off.

H Additional Qualitative Comparisons

This section reports additional qualitative cases for complex camera/sequential motion and VBench-Long prompts.

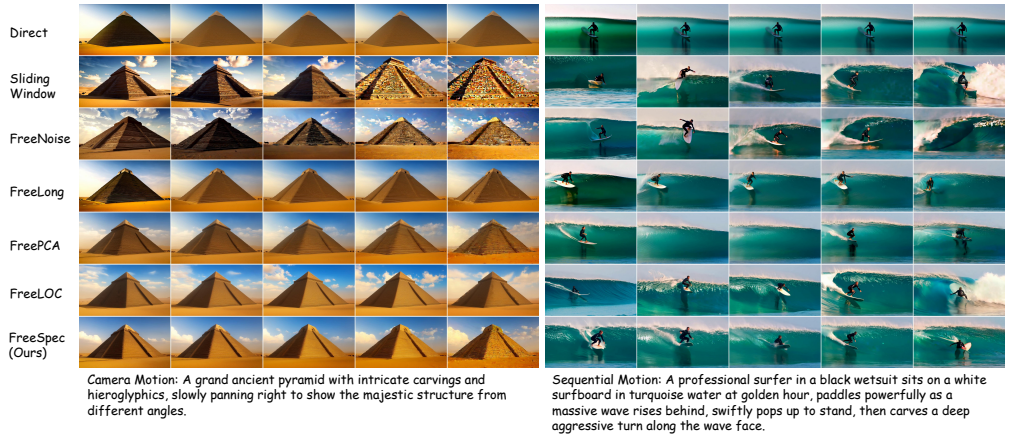

Figure A4: Additional complex-motion examples on Wan2.1 [7]. The results compare camera-motion and sequential-motion generation under $4\times$ the native training length.

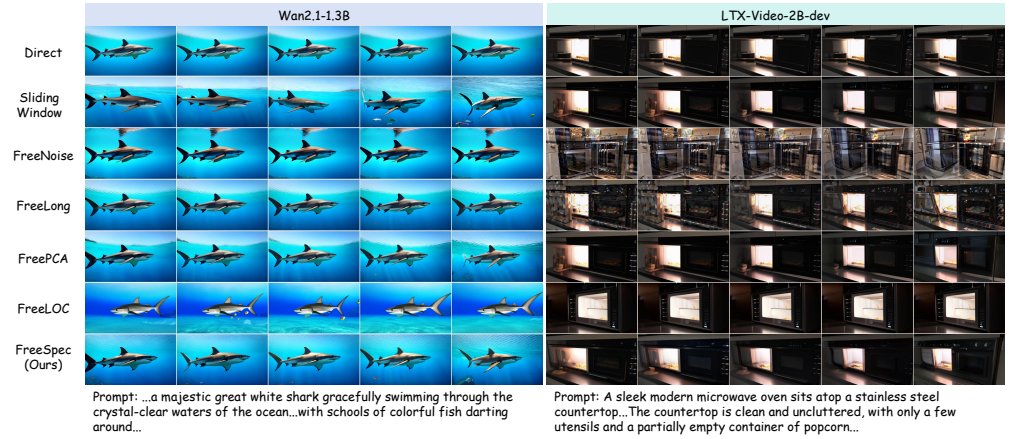

Figure A5: Additional qualitative comparison for long-video generation.

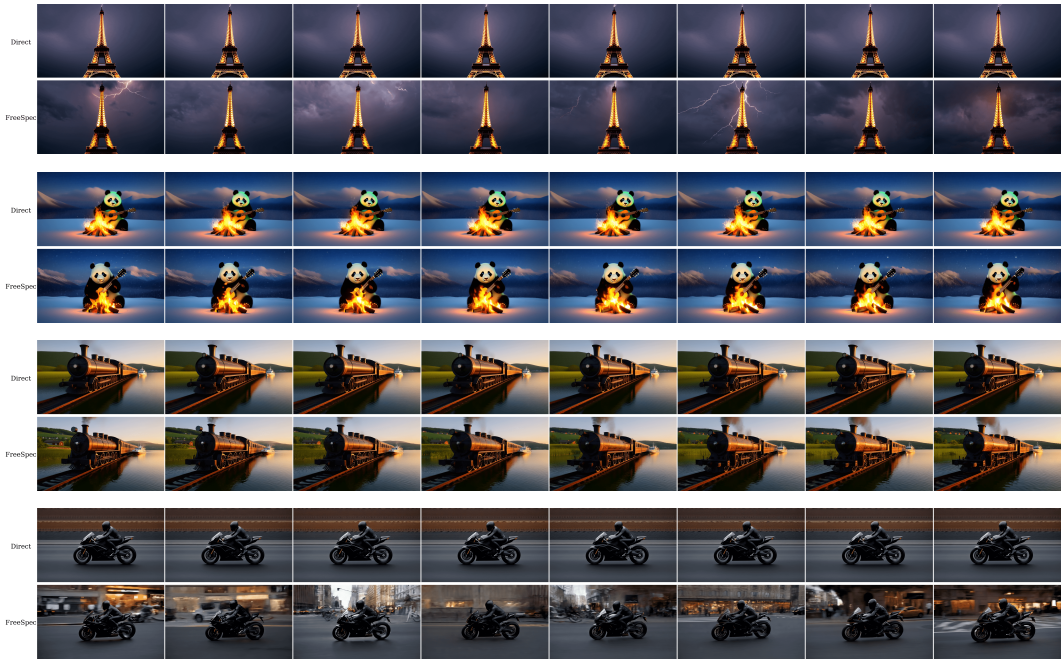

Figure A6: Additional VBench-Long comparisons between Direct and FreeSpec.

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: The abstract and introduction clearly state the main claims: enlarged self-attention windows induce spectral concentration, FreeSpec mitigates this issue through training-free singular-spectrum reconstruction, and experiments on Wan2.1 and LTX-Video validate its effectiveness. The claims are limited to training-free long video generation with pretrained short-video diffusion models and are supported by the analysis and experiments in the paper.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: The paper includes a dedicated Failure Cases and Limitations section. It discusses that FreeSpec inherits the semantic reasoning limitations of the frozen base model and may fail to infer implicit scene transitions in multi-stage actions.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or proofs.

   Guidelines:

   - The answer [N/A] 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: The paper describes the evaluated backbones, compared methods, benchmark prompts, evaluation metrics, generation length, algorithmic components, and major hyperparameters. The appendix provides additional implementation details for reproducing the training-free inference procedure.

   Guidelines:

   - The answer [N/A] 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: The source code is publicly available via the link provided in the abstract.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies that FreeSpec is training-free and thus does not require optimizer or training details. It reports the evaluated prompts, backbones, baselines, metrics, generation length, ablation variants, and key inference hyperparameters.

   Guidelines:

   - The answer [N/A] 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: The paper reports quantitative results on 100 VBench-Long prompts and ablations on 50 prompts, but the main comparison tables do not include error bars or statistical significance tests. The spectral observation analysis reports mean curves with standard error, but this does not cover all main experimental results.

Guidelines:

- The answer [N/A] 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.
- 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: The paper reports inference time for each method and backbone in the main quantitative tables. The implementation details further describe the inference setting and resource requirements needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] 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: The work studies a training-free inference method for long video generation using existing pretrained models and public evaluation protocols. It does not involve private or sensitive data, human-subject experiments, or deceptive deployment. FreeSpec inherits the general risks of generative video models, including misleading synthetic media, impersonation, and copyright-sensitive content. Practical deployment should include provenance tracking, watermarking, content moderation, and responsible-use policies.

   Guidelines:

   - The answer [N/A] 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: The paper discusses the research benefits of training-free long-video extension, as well as risks from deceptive media and disinformation, with provenance tracking, watermarking, and content moderation as possible mitigations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained generative model or a newly collected dataset. FreeSpec is an inference-time method applied to existing video diffusion models, whose release and safety mechanisms are governed by the corresponding base-model providers.

    Guidelines:

    - The answer [N/A] 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: The paper cites the existing pretrained video generation backbones, baselines, and evaluation benchmarks used in the experiments, including Wan2.1, LTX-Video, FreeNoise, FreeLong, FreePCA, and VBench-based evaluation. The experiments are conducted using these existing assets for research comparison.

Guidelines:

- The answer [N/A] 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.
- 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: [N/A]

Justification: The paper does not introduce a new dataset, pretrained model, or benchmark asset. It proposes a training-free inference method evaluated on existing models and benchmarks.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or human-subject experiments. All evaluations are based on automatic metrics and qualitative comparisons.

Guidelines:

- The answer [N/A] 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: [N/A]

    Justification: The paper does not involve research with human subjects and therefore does not require IRB approval or equivalent review.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method. LLMs were used only for writing, editing, or formatting, such use does not affect the core methodology or experimental results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.