SpecLoR: Improving Video Generation with Inference-Time Spectral Lookahead Rectification

Xu Zhang[1], **Yu Lu**[1], **Ruijie Quan**[1], **Zhaozheng Chen**[2], **Bohan Wang**[2], **Yi Yang**[1*]

[1] ReLER, College of Artificial Intelligence, Zhejiang University [2] Huawei Central Research Institute

Project Page: https://xuzhang2.github.io/SpecLoR/

Abstract

Flow Matching has enabled high-fidelity text-to-video generation via latent ODE sampling. However, velocity approximation and numerical discretization errors inevitably accumulate during inference, causing sampling trajectories to drift from the intended generative path. Consequently, generated videos often suffer from severe spatiotemporal inconsistencies. A natural idea is to correct this drift during sampling, but direct intervention in the noisy latent space is challenging: (i) timestep-dependent noise obscures reliable structural cues; (ii) spatial interventions risk disrupting intricate local geometry while incurring heavy computational costs. To address these challenges, we propose Spectral Lookahead Rectification (**SpecLoR**), an inference-time method that bypasses noise via lookahead prediction, and circumvents spatiotemporal entanglement by shifting corrections to the frequency domain. First, during early sampling stages, SpecLoR estimates the clean latent $z_{t,0}$ from the current noisy state and computes its 3D spatiotemporal spectrum. Next, SpecLoR rectifies the amplitude spectrum to align with natural-video spectral statistics. Finally, the corrected state is re-noised back to the current timestep to resume ODE integration. Experiments on Wan2.2 show that SpecLoR reduces physical artifacts and enhances motion coherence across multiple benchmarks with minimal computational overhead (4 additional NFEs).

1 Introduction

Text-to-video (T2V) generation [1–8] has recently achieved remarkable milestones, largely driven by the transition toward the Flow Matching (FM) paradigm [9, 10]. Despite great progress, current T2V models still suffer from spatiotemporal inconsistencies, *e.g.*, duplicated limbs, floating objects, or physically implausible contacts, especially under complex motion and interaction dynamics.

A primary cause of these inconsistencies lies in the *sampling drift* during inference [11–13]. In FM models, video generation proceeds by integrating a learned vector field using a numerical ODE solver. Because this estimated field cannot perfectly match the true continuous transport field, velocity approximation and numerical discretization errors inevitably accumulate across integration steps [12]. Consequently, the sampling trajectory gradually deviates from high-density regions of the data distribution, forcing intermediate latent states off the ideal generative path, as illustrated in Fig. 1a. Existing inference-time methods attempt to mitigate this drift through trajectory exploration or perturbation. Search-based approaches [14–16] rely on rejection sampling or test-time selection to explore optimal sampling paths. However, these methods are often computationally prohibitive for high-dimensional video data. Noise initialization techniques [11, 17] tackle this issue at the trajectory's starting point, but leave the overall ODE drift fundamentally unresolved. Meanwhile, perturbation-based methods [12, 13] delve deeper into the intra-trajectory generation process by periodically injecting stochastic noise to dilute accumulated errors during inference. While mitigating

*Corresponding Author: Yi Yang.

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

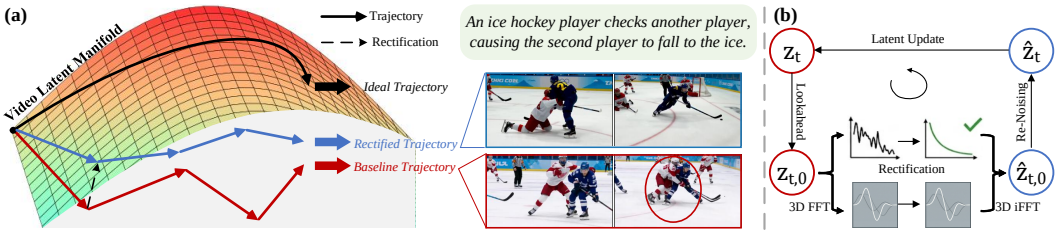

Figure 1: Concept and pipeline of SpecLoR. (a) Accumulated errors drift the Flow Matching trajectory from the ideal trajectory, yielding spatiotemporal inconsistencies. SpecLoR seamlessly anchors it back to a more valid trajectory. (b) At early timesteps, we project the noisy latent z_t to a lookahead state $z_{t,0}$. We decouple its spatiotemporal spectrum, rectifying only the amplitude using empirical video statistics. The corrected state $\hat{z}_{t,0}$ is then re-noised to $\hat{z}_t$, resuming ODE integration.

accumulated errors, such blind perturbations lack a deterministic direction to steer the drifted trajectory back toward the natural video manifold, inadvertently introducing secondary distortions.

Rather than modifying the trajectory directly within the noisy latent space, where timestep-dependent noise severely obscures reliable spatiotemporal structural cues, we bypass this interference by projecting the latent onto a clean lookahead anchor. Crucially, we observe *spectral amplitude distortion* as a correctable manifestation of trajectory drift. Real-world videos exhibit a characteristic spectral statistics, with their ensemble-averaged spatiotemporal amplitude following a $1/f^{\alpha}$ power law [18]. Ideally, the predicted lookahead state should reconstruct the clean video latent and preserve this natural spectral prior. However, velocity approximation and discretization errors during early, high-noise steps cause its amplitude spectrum to deviate from this statistics. Such deviations are often accompanied by visible artifacts such as ghosting, duplicated structures, and unnatural motion. Since these distortions emerge while the global motion structure is still being formed, restoring the natural spectral prior at early stages is particularly effective for correcting trajectory drift.

Motivated by this insight, we introduce **Spectral Lookahead Rectification (SpecLoR)**, an inference-time sampling method that mitigates trajectory drift via spectral amplitude alignment. As illustrated in Fig. 1b, SpecLoR first projects the noisy latent z_t onto its theoretical lookahead anchor, $z_{t,0}$, and extracts its spatiotemporal spectrum. Then, the amplitude spectrum is rectified using this empirical prior while preserving the intricate, geometry-encoding phase [19]. Finally, the corrected latent $\hat{z}_{t,0}$ is re-noised back to the current timestep as $\hat{z}_t$ to resume ODE integration. By restoring the natural spectral magnitude during early sampling stages, SpecLoR acts as a reliable structural anchor. It updates the sampling trajectory using the rectified clean anchor, allowing integration to resume from this corrected state. Comprehensive evaluations on SOTA video generators demonstrate that SpecLoR effectively suppresses structural artifacts and enhances motion coherence across complex dynamics. Specifically, integrating SpecLoR into the open-sourced Wan2.2 framework substantially elevates visual fidelity and spatiotemporal consistency across multiple challenging benchmarks, while incurring minimal computational overhead (only 4 additional NFEs in a 40-step schedule).

Our contributions can be summarized as follows:

- We identify spectral amplitude distortion as a correctable manifestation of sampling drift, providing a novel mechanistic insight into video artifacts within Flow Matching-based video generation.
- We introduce **SpecLoR**, an inference-time method that effectively calibrates the amplitude spectrum of clean lookahead predictions, reintegrating the rectified estimate into the sampling trajectory, substantially enhancing the quality of videos generated by SOTA models (*e.g.*, Wan2.2)
- We reveal that spectral distortions emerge prominently during early sampling stages as global motion characteristics are established. By strategically confining our rectification to this critical window, SpecLoR provides an efficient structural constraint with minimal computational overhead.

2 Related work

2.1 Large-Scale Text-to-Video Generation and Flow Matching

Text-to-Video (T2V) generation has recently witnessed a paradigm shift from iterative Diffusion Models (DMs) [20–22] to continuous-time Flow Matching (FM) [9, 10, 23, 24] paradigms. With

optimal-transport probability paths, FM constructs straighter trajectories than traditional diffusion, enabling faster sampling. Highly scalable Diffusion Transformers [25] have rapidly become the de facto backbone for recent state-of-the-art large video foundation models, including Sora [3], Kling [4], HunyuanVideo [1, 26], and Wan [2]. However, extending FM to high-fidelity, large-scale video synthesis heavily exacerbates a critical vulnerability during inference. The generative process relies on solving an Ordinary Differential Equation (ODE) via discrete numerical integration (*e.g.*, Euler or UniPC [27]). Because predicting the true vector field for high-dimensional, spatiotemporally entangled video data is exceedingly difficult, especially in the high-noise early stages, the network inevitably produces significant velocity approximation errors during inference. As highlighted in recent studies on generative trajectory dynamics [12, 13], these step-by-step errors quickly accumulate into severe trajectory drift, pulling the generated latents off the natural video manifold.

2.2 Inference-Time Sampling and Trajectory Control

To mitigate video generation artifacts, recent works have introduced various inference-time strategies [28–42]. One line of research focuses on guidance modulation; for instance, CFG-Zero* [43] adjusts Classifier-Free Guidance [44] to prevent color over-saturation. While effective for global contrast, it fails to resolve the progressive accumulation of sampling drift. Alternative approaches explore path search [14–16, 45, 46] via rejection sampling or test-time selection; however, these methods are often computationally expensive for high-dimensional video. Similarly, noise initialization techniques [11, 17] solely target the trajectory's initial state, leaving the subsequent ODE drift unmitigated. Another direction involves sampling interventions [13, 12]. Restart Sampling [12], for example, periodically injects noise to re-evaluate problematic segments, yet this stochasticity inadvertently introduces new errors into the trajectory. More recently, test-time optimization approaches such as FlowMo [47] enforce temporal coherence during inference. However, these gradient-dependent methods operate strictly in the spatial domain. Since global motion and fine geometric details are deeply entangled within the spatial latent, direct modifications often unintentionally blur semantic textures or incur prohibitive computational overhead. In contrast, SpecLoR shifts the structural constraint to the frequency domain, where the global amplitude spectrum and instance-specific phase details can be seamlessly decoupled and safely rectified without compromising visual fidelity.

2.3 Spectral Dynamics and Natural Scene Statistics

Analyzing visual signals in the frequency domain provides a mathematically rigorous perspective on generation quality. Natural Scene Statistics demonstrate that real-world videos exhibit a characteristic $1/f^{\alpha}$ power-law decay in their 3D spatiotemporal spectrum [18], which remains preserved within Video Autoencoder latent spaces [48, 22]. While recent image diffusion works [49–51] explore frequency interventions, they primarily reweight 2D spatial features within noise-corrupted intermediate network states. Directly extending these to video fails because ODE trajectory drift fundamentally corrupts the macroscopic 3D energy distribution across frames. Similarly, previous video methods such as FreeInit and FreqPrior [11, 17] use frequency information to improve noise initialization. SpecLoR is fundamentally distinct: rather than modulating noisy spatial features, it operates strictly in the 3D spatiotemporal frequency domain on a noise-free lookahead estimate. By applying empirical natural-video amplitude statistics, SpecLoR rectifies spectral distortion in the lookahead prediction.

3 Spectral Perspective on Trajectory Drift

In Flow Matching (FM), imperfect velocity predictions and numerical discretization inevitably introduce sampling trajectory drift. As errors accumulate, intermediate latent states deviate from the ideal generative path. Visualizing lookahead predictions (Fig. 2a) shows that early structural ambiguities can persist into the final video, as illustrated by the entangled instances at Step 39.

Analyzing this phenomenon from a frequency perspective, we empirically observe that the macroscopic amplitude of the lookahead prediction, which governs global energy distribution, diverges significantly from natural priors during these early stages. These early amplitude deviations coincide with the structural ambiguities visible in the lookahead predictions. This reveals a critical opportunity: mitigating trajectory drift does not necessitate full latent manipulation. In a pilot experiment, we rectified the amplitude spectrum exclusively during early ODE integration (Steps 2-5), leaving the phase untouched. To quantify the resulting dynamics (Fig. 2b), we calculate the correlation of

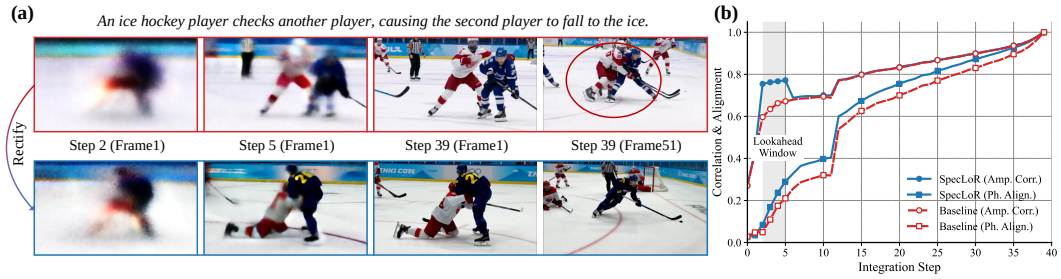

Figure 2: Trajectory drift and spectral rectification. (a) Visualizing intermediate lookahead predictions. Baseline (top): early structural ambiguity at Step 5 degrades into severe physical artifacts by Step 39 (red circles). SpecLoR (bottom) explicitly rectifies amplitude during early stages to establish a coherent global structure. (b) Quantitative tracking of frequency dynamics, measured by correlating intermediate steps against each method's final latent. Amplitude rectification accelerates earlier phase convergence, fundamentally shifting the trajectory toward a more valid endpoint.

intermediate latents against the final generated output at Step 40 of each respective method. By Step 40, the learned noise-to-video transformation has produced a representation that more closely reflects real-video features than the early lookahead predictions. Our objective is to bring these features into the early lookahead prediction, reducing early deviation so that subsequent sampling can build on a more coherent structure. Since a single text condition maps to a highly multimodal distribution of valid videos, we use each method's own final latent to track the relative stabilization of the generative path. The curves reveal that this targeted amplitude intervention implicitly accelerates phase alignment towards this stabilized target, surpassing the baseline during the intervention window and maintaining this advantage. Crucially, this early constraint shifts the trajectory, enabling the model to arrive at a distinctly different and structurally superior endpoint (Fig. 2a). Appendix §C.1 provides additional validation using real-video references.

4 Method

The central challenge in Flow Matching-based video generation is the accumulation of numerical and predictive errors, which inevitably forces the sampling trajectory off the natural video manifold. To address this, we introduce Spectral Lookahead Rectification (SpecLoR), a practical, highly efficient inference method. Classical signal processing theory [19] establishes that the phase spectrum encodes delicate geometric details and is highly vulnerable to arbitrary interventions, which easily introduce severe structural artifacts (as empirically verified in §5.3). Concurrently, as quantitatively demonstrated by our spectral dynamics analysis (Fig. 2b), rectification of the macroscopic energy distribution (amplitude) is associated with earlier phase alignment. Grounded in this insight, our method mitigates trajectory drift via a structured four-stage pipeline, as illustrated in Fig. 3.

4.1 Lookahead Projection

Diagnosing trajectory drift directly within the intermediate noisy latent space z_t is intractable due to the dominance of timestep-dependent noise. To bypass this, we utilize the current predicted vector field $v_\theta(z_t, t)$ to project the state into a noise-free space. As shown in Fig. 3 (Stage 1), we estimate the clean latent $z_{t,0}$ via a single Euler step:

$$z_{t,0} = z_t - t \cdot v_\theta(z_t, t). \tag{1}$$

This lookahead prediction $z_{t,0}$ acts as a diagnostic window, revealing early structural ambiguities before they solidify into physical artifacts. We explicitly constrain this operation to a specific "Lookahead Window" during the early high-noise regime (*e.g.*, steps 2-5 in a 40-step schedule).

4.2 Frequency Decoupling

Directly modifying the spatial representation of $z_{t,0}$ is challenging because global motion structures and fine geometric details are tightly coupled. Any spatial intervention intended to correct this trajectory drift often unintentionally blurs semantic textures. To overcome this entanglement, we shift

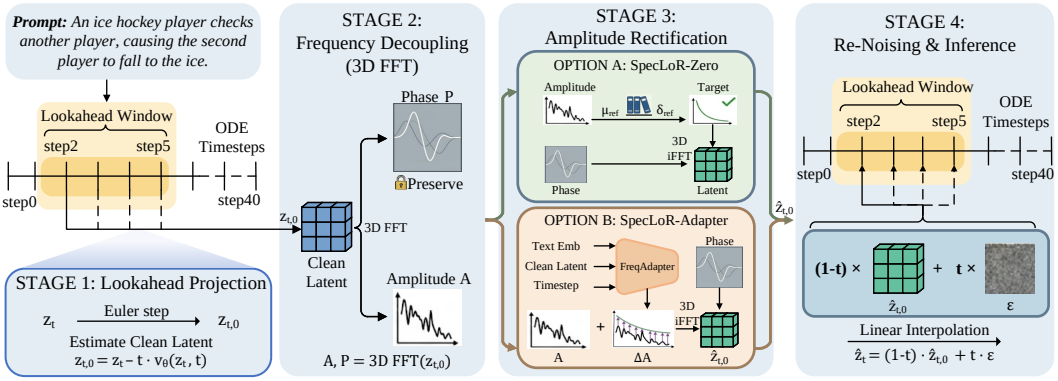

Figure 3: Pipeline of the proposed SpecLoR method. In Stage 1, the intermediate noisy latent z_t is projected to a clean lookahead state $z_{t,0}$. Stage 2 explicitly decouples the spatiotemporal spectrum into amplitude and phase via 3D FFT. To correct trajectory drift, the corrupted amplitude is rectified using either a training-free global prior (Option A) or a context-aware adapter (Option B), while the detail-preserving phase is strictly locked. Finally, in Stage 4, the rectified spectrum is inverted and re-noised back to $\hat{z}_t$, providing a corrected state for the subsequent ODE integration.

our intervention to the frequency domain (Fig. 3, Stage 2). Specifically, we apply a 3D Fast Fourier Transform (FFT) across the temporal and spatial dimensions (T, H, W) of $z_{t,0}$. This transformation explicitly decouples the macroscopic energy (Amplitude $\mathcal{A}$), which is susceptible to trajectory drift yet safe to rectify, from the highly vulnerable, geometry-encoding phase (Phase $\mathcal{P}$). By isolating and strictly preserving the phase $\mathcal{P}$, we establish a decoupled representation where the amplitude can be corrected without compromising instance-specific details or visual fidelity.

4.3 Amplitude Rectification

SpecLoR-Zero (Global Prior). According to Natural Scene Statistics [18], real-world videos exhibit a $1/f^{\alpha}$ power-law spectral decay. Early lookahead predictions exhibit deviations from this statistical pattern. To rectify this, as depicted in Fig. 3 (Option A), we introduce an asymmetric trust region derived from an empirical natural-video prior $(\mu_{ref}, \sigma_{ref})$ precomputed from roughly 2,000 randomly sampled real videos, with statistics computed at each channel and frequency. We bound the current amplitude $\mathcal{A}_{curr}$ within this natural variance envelope to obtain the target amplitude $\mathcal{A}_{tgt}$:

$$\mathcal{A}_{tgt} = \text{clip}(\mathcal{A}_{curr},\ \gamma \cdot \mu_{ref},\ \mu_{ref} + k\sigma_{ref}), \tag{2}$$

where γ (set to 0.6) and k (set to 3) denote the lower and upper margin factors, respectively. We then apply a weighted update to obtain the final rectified amplitude:

$$\mathcal{A}_{rect} = \mathcal{A}_{curr} + \lambda \cdot (\mathcal{A}_{tgt} - \mathcal{A}_{curr}), \tag{3}$$

where λ controls the intervention strength (empirically set to 0.5). This overhead-free structural anchor effectively stabilizes the trajectory, suppressing physically implausible artifacts.

SpecLoR-Adapter (Context-Aware Refinement). While the global prior provides a robust foundation, ideal energy distributions vary across motion dynamics. To capture these instance-specific variations, we introduce the SpecLoR-Adapter (Fig. 3, Option B).

Constructed with Diffusion Transformer (DiT) blocks, the SpecLoR-Adapter leverages the predicted lookahead latent $z_{t,0}$, timestep embeddings t_{emb}, and text conditions to estimate an instance-specific residual amplitude map $\Delta\mathcal{A}$. During offline training on natural videos, we simulate the Flow Matching forward process on a ground-truth latent z_0 (with exact amplitude $\mathcal{A}_{gt}$) to sample an early-stage noisy latent z_t. After deriving the current lookahead amplitude $\mathcal{A}_{curr}$ from the predicted $z_{t,0}$, we enforce nonnegative amplitudes by defining $\mathcal{A}_{rect} = \max(0, \mathcal{A}_{curr} + \Delta\mathcal{A})$, where the maximum is applied element-wise. The adapter is then optimized via a Mean Squared Error (MSE) loss in the logarithmic magnitude space:

$$\mathcal{L} = \|\log(1 + \mathcal{A}_{rect}) - \log(1 + \mathcal{A}_{gt})\|_2^2 . \tag{4}$$

This log transformation compresses the spectrum's massive dynamic range caused by power-law decay. It prevents low-frequency dominance from overwhelming the loss, forcing the adapter to capture fragile high-frequency structural details while stabilizing training gradients.

Crucially, the base video model remains strictly frozen throughout training. At inference, we use the same nonnegative $\mathcal{A}_{rect}$ to reconstruct the rectified clean latent.

4.4 Re-Noising & Inference

Once the amplitude is rectified, we combine $\mathcal{A}_{rect}$ with the preserved phase $\mathcal{P}$. A 3D Inverse FFT (iFFT) projects them back into the latent space, yielding a purified, structurally rectified clean anchor $\hat{z}_{t,0}$. To integrate this corrected state back into the ODE solver (Fig. 3, Stage 4), we re-noise $\hat{z}_{t,0}$ back to the exact current timestep t via linear interpolation with a noise term ϵ:

$$\hat{z}_t = (1-t) \cdot \hat{z}_{t,0} + t \cdot \epsilon. \tag{5}$$

Deterministic Continuation: By setting $\epsilon = \epsilon_{init}$, we reuse the initial noise map. For a fixed seed, this makes the reconstruction deterministic and avoids introducing fresh stochastic noise. Keeping the same noise endpoint supports controlled comparisons between reconstruction with and without amplitude rectification, as evaluated in Table 2.

Stochastic Contraction: By setting $\epsilon \sim \mathcal{N}(0, I)$, we inject freshly sampled stochastic noise. Drawing on the error-contraction properties of generative processes [12], introducing stochasticity helps contract accumulated ODE discretization errors. Since SpecLoR already anchors the rectified macroscopic physical trajectory, this fresh noise optimizes local details without causing spatiotemporal collapse, yielding a globally more physically coherent video distribution.

5 Experiments

To comprehensively evaluate the effectiveness of the proposed SpecLoR, we conduct extensive experiments focusing on motion coherence, overarching visual quality, and computational efficiency.

5.1 Experiment Setup

Benchmarks. We meticulously evaluate our method across two rigorous benchmarks tailored for physical and temporal stability: (1) **Dynamic-Bench** [13] to assess structural stability under challenging, multi-object interactions and complex human motions; and (2) **VideoJAM-Bench** [52] to target the joint appearance-motion coherence of generated dynamics.

Evaluation Metrics. To ensure a multi-dimensional assessment, we categorize our metrics as follows:

- *Visual & Physical Integrity:* Evaluated via the Visual Quality, Physical Consistency, and T2V Alignment scores from VideoScore2 [53], alongside Consistency, defined as the arithmetic mean of the Subject Consistency and Background Consistency scores from VBench [54].
- *Temporal Dynamics:* Assessed using Motion Smoothness and Dynamic Degree from VBench.
- *Computational Efficiency:* Strictly tracked via the Number of Function Evaluations (NFE).

Implementation Details. Built upon `Wan2.2-A14B T2V` [2] (default 40 NFEs), SpecLoR operates exclusively during the early "Lookahead Window" (Steps 2-5). This targeted, inference-only intervention inherently bounds computational overhead, requiring only 44 NFEs. Experiments are conducted on a 910B NPU with 60GB of memory, taking approximately 20 minutes to generate a single video. Additional backbone evaluations appear in Appendices §D.3.1 and §D.3.2.

5.2 Evaluation on Complex Dynamics

Generating complex dynamics often exposes the fragility of the standard sampling trajectory. We evaluate the performance of SpecLoR against existing methods across the following four dimensions:
Quantitative Comparison. As summarized in Table 1, both variants of SpecLoR improve physical-consistency and visual-quality scores over the base model across Dynamic-Bench and VideoJAM-Bench. On VideoJAM-Bench, SpecLoR-Adapter achieves the highest score in T2V Alignment (4.126), while SpecLoR-Zero achieves the highest Physical Consistency (3.448) and Visual Quality (3.939). On Dynamic-Bench, SpecLoR-Zero establishes the top T2V Alignment (3.651), while SpecLoR-Adapter achieves the highest Visual Quality (3.864). Compared to the baseline, our method effectively enhances Physical Consistency while maintaining highly competitive Motion Smoothness (*e.g.*, reaching 0.980 on Dynamic-Bench). This indicates that amplitude rectification preserves smooth temporal progression of the video.

Table 1: Quantitative comparison. Best scores are in **bold** and second best are underlined.

Method	VideoScore2			VBench		Efficiency	
	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑	NFE↓	Time↓
Results on VideoJAM-Bench [52]							
Wan2.2 T2V	4.062	3.237	3.873	0.935	0.972	40	
+ NFE×1.5	4.046	3.260	3.834	0.935	0.974	60	1.5×
+ FlowMo [47]	4.044	3.179	3.805	0.939	0.974	40	1.7×
+ SRVS [13]	4.067	3.341	3.926	**0.943**	**0.979**	60	1.5×
+ SpecLoR-Zero (Ours)	4.118$_{\pm 0.013}$	**3.448**$_{\pm 0.018}$	**3.939**$_{\pm 0.017}$	0.939$_{\pm 0.002}$	0.977$_{\pm 0.004}$	44	1.1×
+ SpecLoR-Adapter (Ours)	**4.126**$_{\pm 0.011}$	3.441$_{\pm 0.019}$	3.927$_{\pm 0.014}$	0.940$_{\pm 0.003}$	0.977$_{\pm 0.002}$	44	1.1×
Results on Dynamic-Bench [13]							
Wan2.2 T2V	3.605	2.687	3.745	0.909	0.974	40	
+ NFE×1.5	3.541	2.898	3.688	0.912	0.974	60	1.5×
+ FlowMo [47]	3.585	**3.086**	3.745	0.911	0.975	40	1.7×
+ SRVS [13]	3.584	2.993	3.800	**0.917**	**0.981**	60	1.5×
+ SpecLoR-Zero (Ours)	**3.651**$_{\pm 0.012}$	2.993$_{\pm 0.012}$	3.810$_{\pm 0.014}$	0.915$_{\pm 0.005}$	0.980$_{\pm 0.003}$	44	1.1×
+ SpecLoR-Adapter (Ours)	3.588$_{\pm 0.014}$	3.058$_{\pm 0.021}$	**3.864**$_{\pm 0.013}$	0.914$_{\pm 0.004}$	0.980$_{\pm 0.002}$	44	1.1×

Computational Efficiency. SpecLoR is designed as an exceptionally lightweight, inference-time intervention. As shown in the Efficiency columns of Table 1, SpecLoR requires only 44 NFEs, introducing a minimal overhead (1.1× Time) over the standard 40-step baseline. This provides a highly favorable performance-compute trade-off compared to simply scaling up the ODE solver steps (NFE×1.5), which demands 60 NFEs and a 1.5× time penalty. Furthermore, SpecLoR maintains a decisive efficiency advantage when benchmarked against other test-time techniques like SRVS (60 NFEs, 1.5× Time) and FlowMo (1.7× Time). At the same 44-NFE budget and seed, Adapter exceeds SRVS on all five metrics on both benchmarks; Zero exceeds it on four VideoJAM metrics and all five Dynamic-Bench metrics. Both variants also achieve higher T2V Alignment and Physical Consistency than FreeInit, FreqPrior, CFG-Zero*, and Enhance-A-Video in the additional VideoJAM comparison. See Appendix §B.2 for the memory overhead and Appendix §D.1 for SRVS at 44 NFEs and additional baseline comparisons.

User Study. We conduct a blind human preference study on VideoJAM-Bench using SpecLoR-Adapter with Stochastic Contraction, which samples fresh Gaussian noise during re-noising. Five evaluators independently assess the same 128 video pairs, giving 640 judgments per criterion. Annotators evaluate randomized video pairs (the Wan2.2 baseline vs. SpecLoR) across three criteria: Text Alignment, Motion Coherence, and Video Quality. As illustrated in Fig. 4, SpecLoR consistently outperforms the baseline. Specifically, our method receives substantially higher preference votes in Video Quality (43.0% vs. 17.9%) and Motion Coherence (39.8% vs. 16.4%). Furthermore, while the robust baseline model already exhibits excellent capability in Text Alignment, resulting in a high tie rate of 68.8%, SpecLoR still maintains a noticeable edge (20.3% vs. 10.9%). These results confirm that SpecLoR successfully translates into visually superior and motion-coherent dynamics, while preserving and even enhancing the base model's strong prompt adherence. All five evaluators favor SpecLoR on each criterion. Appendix §D.4 reports study size, evaluator agreement, and statistical analysis.

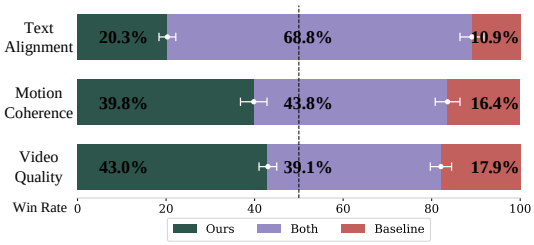

Figure 4: User study on VideoJAM-Bench. Error bars show standard deviation across five evaluators.

Qualitative Results. Visual inspections (as shown in Fig. 5) corroborate our quantitative findings. While baseline trajectories and extended sampling approaches occasionally struggle with geometric distortions or semantic ambiguities under intense motion, SpecLoR yields clearer structural boundaries and fewer physically implausible artifacts. This explicitly aligns with the improved Physical Consistency metric, providing further evidence that SpecLoR effectively mitigates trajectory drift.

5.3 Diagnostic Experiment

We validate SpecLoR on Wan2.2 using two mentioned benchmarks. To isolate intervention effects, all diagnostics on Dynamic-Bench (except Table 4) strictly employ Deterministic Continuation.

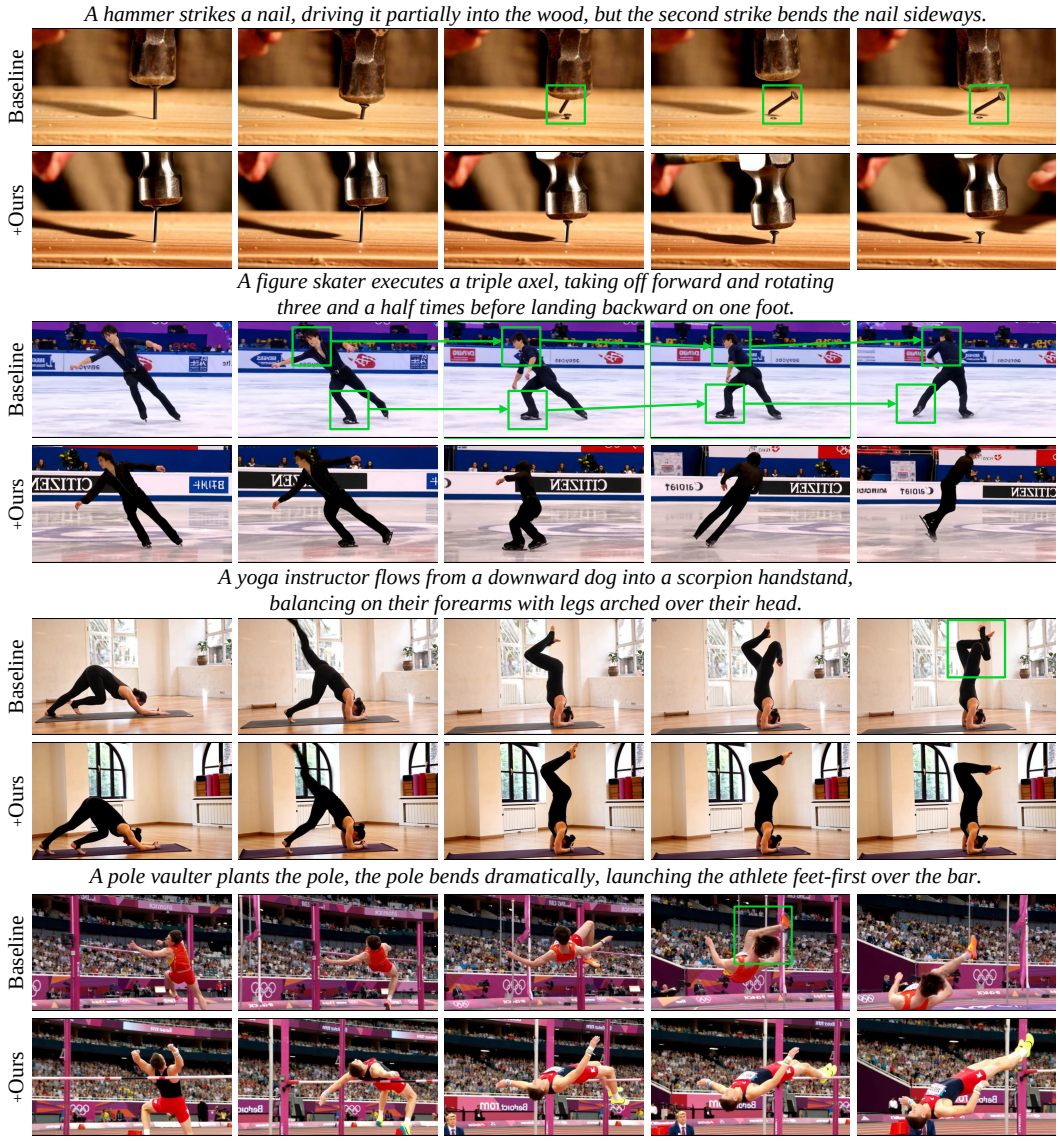

Figure 5: Qualitative comparison of spatially consistent video generation against Wan2.2-A14B T2V. Artifacts and structural errors are indicated by green bounding boxes.

Table 2: Contribution of amplitude rectification on VideoJAM-Bench. The last two rows use the same lookahead and reconstruction procedure; only the full method applies amplitude rectification.

Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑
Baseline	4.062	3.237	3.873	0.935	0.972
Lookahead + reconstruction	4.068	3.316	3.879	0.936	0.973
SpecLoR-Zero	**4.113**	**3.432**	**3.909**	**0.937**	**0.974**

Contribution of Amplitude Rectification. Table 2 isolates the contribution of amplitude rectification by comparing the baseline, lookahead and reconstruction alone, and full SpecLoR-Zero on VideoJAM-Bench. The latter two differ only in whether the predicted amplitude is rectified. Lookahead and reconstruction raise Physical Consistency from 3.237 to 3.316; adding amplitude rectification increases it further to 3.432 and improves all five metrics over this control. The larger incremental gain from rectification highlights its central contribution to physical consistency, alongside improved visual quality and text alignment.

Intervention Target: Amplitude vs. Phase. We investigate what happens if we intervene on the phase instead of the amplitude. We conducted experiments rectifying only the phase using global

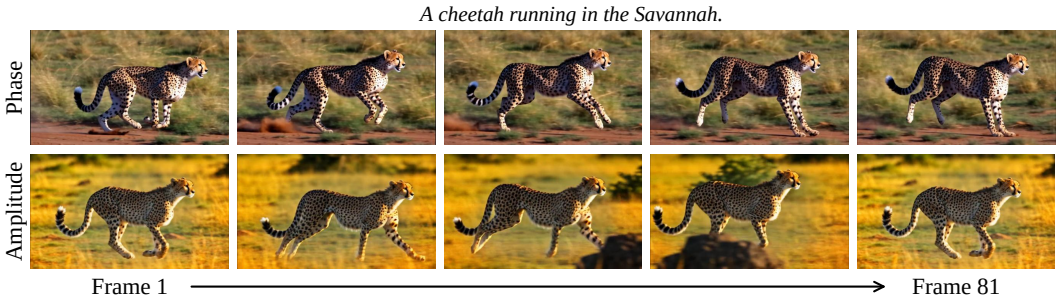

Figure 6: Diagnostic comparison of intervention targets (Phase *vs.* Amplitude).

Table 3: Quantitative evaluation of intervention window timing on VideoJAM-Bench. Early intervention (Steps 2–5) yields higher consistency and motion smoothness than the later windows tested.

Method	Consistency↑	Motion Smooth.↑	Dynamic Degree↑	NFE↓	Time↓
Wan2.2 T2V	0.935	0.972	**0.875**	40	
+ SpecLoR-Zero (step 6-9)	0.935	0.971	0.844	44	1.1×
+ SpecLoR-Zero (step 10-13)	0.933	0.971	0.844	44	1.1×
+ SpecLoR-Zero (step 14-17)	0.932	0.970	0.856	44	1.1×
+ SpecLoR-Zero (step 2-5)	0.939	0.973	0.859	44	1.1×
+ SpecLoR-Adapter (step 2-5)	**0.942**	**0.978**	0.783	44	1.1×

Table 4: Ablation on re-noising stochasticity.

Setting	Consist.↑	Smooth.↑	Dyn. Deg.↑
Zero (Det.)	0.909	0.974	**0.916**
Zero (Stoch.)	**0.912**	**0.977**	0.850
Adapt. (Det.)	0.910	0.973	**0.875**
Adapt. (Stoch.)	**0.914**	**0.979**	0.800

Table 5: Ablation on modulation strength (λ).

Method	Consist.↑	Smooth.↑	Dyn. Deg.↑
$\lambda = 1.0$	0.908	0.973	0.908
$\lambda = 0.8$	0.909	0.973	0.883
$\lambda = 0.5$	0.909	**0.974**	**0.916**
$\lambda = 0.2$	**0.910**	**0.974**	0.900

natural statistics. Rather than merely generating spatial artifacts, altering the instance-specific phase directly led to a fundamental breakdown of the underlying motion structure. As demonstrated in Fig.6, arbitrarily altering the phase disrupts delicate temporal alignment, causing severe dynamic anomalies (*e.g.*, forcing fast-moving actions into disjointed, extremely slow-motion sequences). Because phase inherently encodes spatiotemporal displacement, its disruption primarily corrupts motion coherence without adding typical spatial noise. These observations support preserving phase while rectifying amplitude during early sampling.

Lookahead Window. Table 3 compares four-step intervention windows on VideoJAM-Bench at the same 44-NFE budget. Zero at Steps 2–5 improves Consistency and Motion Smoothness over the baseline; moving the window to Steps 6–9, 10–13, or 14–17 reduces both scores. Early correction also retains more motion than the later windows, with Dynamic Degree of 0.859 versus 0.844–0.856. Adapter in the early window reaches the highest Consistency and Motion Smoothness in this ablation. These results connect the early spectral behavior in §3 to the chosen intervention window. Appendix §C.4 provides the extended timing analysis.

The Role of Re-Noising Stochasticity. We investigate the impact of the noise term ϵ during re-noising (Table4). Strict Deterministic Continuation ($\epsilon = \epsilon_{init}$) retains a higher Dynamic Degree while facilitating precisely controlled visual comparisons. Conversely, Stochastic Contraction ($\epsilon \sim \mathcal{N}(0, I)$) injects freshly sampled noise. In this setting, stochastic continuation yields superior Consistency and Motion Smoothness. However, this reveals a fundamental consistency-dynamism trade-off, a phenomenon similarly observed in perturbation-based methods like SRVS [13]. The observed trade-off favors global spatiotemporal stability over peak dynamic intensity.

Modulation Strength (λ) Dynamics. As detailed in Table5, we evaluated the intervention performance using $\lambda \in \{0.2, 0.5, 0.8, 1.0\}$. While overall quantitative differences are bounded, extreme values introduce specific qualitative failure modes. A weak intervention ($\lambda = 0.2$) yields a lower Dynamic Degree than $\lambda = 0.5$. Conversely, a highly aggressive replacement ($\lambda = 1.0$) can cause the

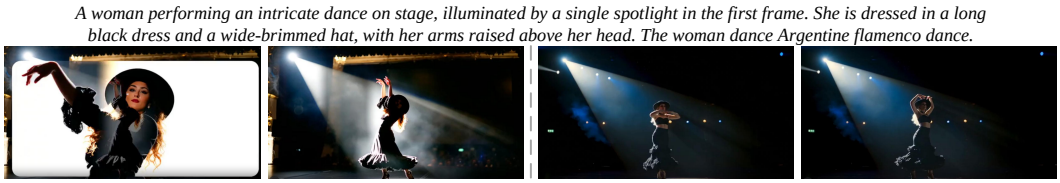

Figure 7: Impact of modulation strength (λ). An aggressive intervention with $\lambda = 1.0$ (left) can produce temporal artifacts, sometimes visually desynchronizing the initial frame from the sequence.

initial frame to become visually uncoordinated with the rest of the sequence (Fig. 7). Setting $\lambda = 0.5$ provides a firm energy anchor while retaining part of the sample-specific amplitude spectrum.

6 Discussion and Conclusion

Limitations. While SpecLoR stabilizes structure by preserving the phase spectrum, this strict adherence can be a double-edged sword. If initial phase predictions are severely corrupted during the earliest ODE steps, locking the phase might trap extreme non-rigid transformations into unnatural states. Additionally, computing a global 3D spatiotemporal FFT introduces memory overhead that scales with spatial resolution and sequence length, and this formulation binds the pre-computed statistical prior to a fixed shape, necessitating a complete re-extraction if the resolution changes.

Broader Impacts. By significantly enhancing the visual realism and spatiotemporal consistency of text-to-video models, our framework inadvertently amplifies dual-use risks. The capability to generate structurally coherent, photorealistic dynamics could be misused to create convincing deepfakes or misinformation. Responsible deployment of such enhanced foundational models necessitates the concurrent development of robust safeguards, including resilient latent watermarking [55, 56] and video forgery detection [57], to mitigate potential societal harms.

Conclusion. In this work, we address a fundamental bottleneck in Flow Matching-based video generation: the unchecked accumulation of trajectory drift during ODE integration, which irrecoverably corrupts spatiotemporal and structural coherence. To overcome this, we introduce Spectral Lookahead Rectification (SpecLoR), an efficient inference paradigm that shifts trajectory correction to the frequency domain. By diagnosing drift via lookahead projection, our method strategically anchors the macroscopic amplitude to match natural video priors during the earliest inference steps. Our controlled experiments show that frequency-specific amplitude correction improves motion coherence beyond lookahead reconstruction and smoothing. By mitigating trajectory drift with a minimal overhead of merely 4 additional NFEs, SpecLoR improves the cost-performance trade-off in inference-time optimization. Ultimately, this plug-and-play method boosts large-scale video foundation models, advancing reliable and motion-coherent spatiotemporal generation.

Acknowledgement. This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM103).

References

[1] 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.

[2] 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.

[3] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[4] Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-Omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

[5] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *Int. Conf. Learn. Represent.*, 2023.

[6] Yifu Zhang, Hao Yang, Yuqi Zhang, Yifei Hu, Fengda Zhu, Chuang Lin, Xiaofeng Mei, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Waver: Wave your way to lifelike video generation. *arXiv preprint arXiv:2508.15761*, 2025.

[7] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[8] Team Seedance. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[9] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Int. Conf. Learn. Represent.*, 2023.

[10] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Int. Conf. Learn. Represent.*, 2023.

[11] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging initialization gap in video diffusion models. In *Eur. Conf. Comput. Vis.*, pages 378–394. Springer, 2024.

[12] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *Adv. Neural Inform. Process. Syst.*, pages 76806–76838, 2023.

[13] Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. *arXiv preprint arXiv:2601.18577*, 2026.

[14] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Int. Conf. Mach. Learn.*, 2024.

[15] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. In *Adv. Neural Inform. Process. Syst.*, 2025.

[16] Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. *arXiv preprint arXiv:2505.17618*, 2025.

[17] Yunlong Yuan, Yuanfan Guo, Chunwei Wang, Wei Zhang, Hang Xu, and Li Zhang. FreqPrior: Improving video diffusion models with frequency filtering gaussian noise. In *Int. Conf. Learn. Represent.*, 2025.

[18] Dawei W Dong and Joseph J Atick. Statistics of natural time-varying images. *Network: computation in neural systems*, 6(3):345, 1995.

[19] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 6840–6851, 2020.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.

[22] 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.

[23] Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Int. Conf. Learn. Represent.*, 2023.

[24] Maojiang Su, Mingcheng Lu, Jerry Yao-Chieh Hu, Shang Wu, Zhao Song, Alex Reneau, and Han Liu. A theoretical analysis of discrete flow matching generative models. *arXiv preprint arXiv:2509.22623*, 2025.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Int. Conf. Comput. Vis.*, pages 4195–4205, 2023.

[26] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[27] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In *Adv. Neural Inform. Process. Syst.*, volume 36, pages 49842–49869, 2023.

[28] Harold Haodong Chen, Haojian Huang, Xianfeng Wu, Yexin Liu, Yajing Bai, Wen-Jie Shu, Harry Yang, and Ser-Nam Lim. Temporal regularization makes your video generator stronger. *arXiv preprint arXiv:2503.15417*, 2025.

[29] Xueji Fang, Liyuan Ma, Zhiyang Chen, Mingyuan Zhou, and Guo-jun Qi. InfLVG: Reinforce inference-time consistent long video generation with grpo. *arXiv preprint arXiv:2505.17574*, 2025.

[30] Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-time scaling for video generation. In *Int. Conf. Comput. Vis.*, pages 18671–18681, 2025.

[31] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[32] Yang Luo, Xuanlei Zhao, Mengzhao Chen, Kaipeng Zhang, Wenqi Shao, Kai Wang, Zhangyang Wang, and Yang You. Enhance-A-Video: Better generated video for free. *arXiv preprint arXiv:2502.07508*, 2025.

[33] Hyelin Nam, Jaemin Kim, Dohun Lee, and Jong Chul Ye. Optical-flow guided prompt optimization for coherent video generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7837–7846, 2025.

[34] 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 *Int. Conf. Learn. Represent.*, 2024.

[35] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. FreeLong: Training-free long video generation with spectralblend temporal attention. In *Adv. Neural Inform. Process. Syst.*, pages 131434–131455, 2024.

[36] Zhuoling Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. LongDiff: Training-free long video generation in one go. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17789–17798, 2025.

[37] Dohun Lee, Bryan Sangwoo Kim, Geon Yeong Park, and Jong Chul Ye. VideoGuide: Improving video diffusion models without training through a teacher's guide. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2599–2608, 2025.

[38] Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C. Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *Int. Conf. Learn. Represent.*, 2024.

[39] Mariam Hassan, Bastien Van Delft, Wuyang Li, and Alexandre Alahi. Factorized video generation: Decoupling scene construction and temporal synthesis in text-to-video diffusion models. *arXiv preprint arXiv:2512.16371*, 2025.

[40] Yang Fei, George Stoica, Jingyuan Liu, Qifeng Chen, Ranjay Krishna, Xiaojuan Wang, and Benlin Liu. Structure from tracking: Distilling structure-preserving motion for video generation. *arXiv preprint arXiv:2512.11792*, 2025.

[41] Ahmet Berke Gokmen, Yigit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. RoPECraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers. *arXiv preprint arXiv:2505.13344*, 2025.

[42] Chenhui Zhu, Yilu Wu, Shuai Wang, Gangshan Wu, and Limin Wang. MotionRAG: Motion retrieval-augmented image-to-video generation. *arXiv preprint arXiv:2509.26391*, 2025.

[43] Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. CFG-Zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

[44] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[45] Byeonghu Na, Yeongmin Kim, Minsang Park, Donghyeok Shin, Wanmo Kang, and Il Chul Moon. Diffusion rejection sampling. In *Int. Conf. Mach. Learn.*, volume 235, pages 37097–37121, 2024.

[46] Vignav Ramesh and Morteza Mardani. Test-time scaling of diffusion models via noise trajectory search. In *Adv. Neural Inform. Process. Syst.*, 2025.

[47] Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. FlowMo: Variance-based flow guidance for coherent motion in video generation. *arXiv preprint arXiv:2506.01144*, 2025.

[48] Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Improved video vae for latent video diffusion model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18124–18133, 2025.

[49] Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Yichen Zheng, Bineng Zhong, Chongyi Li, and Liqiang Nie. Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25378–25388, 2024.

[50] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion u-net. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4733–4743, 2024.

[51] Haosen Yang, Adrian Bulat, Isma Hadji, Hai X Pham, Xiatian Zhu, Georgios Tzimiropoulos, and Brais Martinez. FAM Diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2459–2468, 2025.

[52] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

[53] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. VideoScore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

[54] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21807–21818, 2024.

[55] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian Shading: Provable performance-lossless image watermarking for diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12162–12171, 2024.

[56] Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Mingyi Hong, and Jie Ding. RAW: A robust and agile plug-and-play watermark framework for ai-generated images with provable guarantees. In *Adv. Neural Inform. Process. Syst.*, volume 37, pages 132077–132105, 2024.

[57] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *Adv. Neural Inform. Process. Syst.*, volume 37, pages 122054–122077, 2024.

[58] Runze Zhang, Guoguang Du, Xiaochuan Li, Qi Jia, Liang Jin, Lu Liu, Jingjing Wang, Cong Xu, Zhenhua Guo, Yaqian Zhao, Xiaoli Gong, Rengang Li, and Baoyu Fan. DropletVideo: A dataset and approach to explore integral spatio-temporal consistent video generation. In *Int. Conf. Comput. Vis.*, pages 15583–15593, 2025.

[59] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[60] 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.

[61] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Overview of the Supplementary Material

The supplementary material is organized into four sections:

- **§A: Methodological Analysis.** Verification of the spectral prior, prior statistics and rectifier choices, fixed-endpoint reconstruction, and discussion of related trajectory-control methods.
- **§B: Implementation and Reproducibility.** Data, inference pseudocode, prior extraction, adapter training, memory overhead, evaluation settings, and asset licenses.
- **§C: Controlled Experiments and Ablations.** Real-video reference tests, amplitude smoothing and shuffled-prior controls, and sensitivity to the intervention window.
- **§D: Extended Evaluation.** Additional baseline comparisons, MovieGen-Bench results, generalization to HunyuanVideo-1.5 and LTX-Video, human preference evaluation, and failure cases.

A Methodological Analysis

A.1 Empirical Verification of Spatiotemporal Power-Law

According to Natural Scene Statistics [18], real-world visual signals naturally exhibit a power-law decay in their amplitude spectrum. To empirically verify that this fundamental property is preserved within the highly compressed 3D latent space of modern video autoencoders, we aggregate the amplitude spectrum across 2,000 clean videos. Specifically, after computing the 3D FFT, we map the 3D Cartesian frequency coordinates (u, v, w) to a 1D radial frequency $f = \sqrt{u^2 + v^2 + w^2}$. Here, (u, v, w) explicitly represent the normalized digital frequencies obtained via FFT shifting, which are strictly bounded by the Nyquist limit $[-0.5, 0.5]$ across all dimensions. By averaging the amplitude values within discretized radial frequency bins, we collapse the 4D amplitude tensor $\mathcal{A} \in \mathbb{R}^{C\times T\times H\times W}$ into a simplified 1D radial amplitude profile $\mathcal{A}(f)$. As illustrated in Fig. S1, when plotted on a log-log scale, the extracted global prior approximately follows a linear trend, consistent with the relation $\mathcal{A}(f) \propto 1/f^{\alpha}$. Our linear regression yields an empirical spatiotemporal scaling exponent of $\alpha \simeq 1.66$.

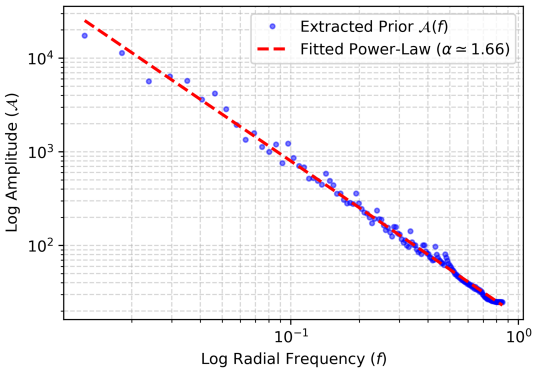

Figure S1: Empirical verification of the power-law prior in the 3D VAE latent space.

A.2 Prior Statistics and the Zero–Adapter Trade-off

Stability of the empirical prior. The global prior uses the InternVid-10M-FLT subset described in Appendix §B.1. We randomly divide the approximately 2,000 clips into two disjoint halves and extract statistics independently. The correlations between the two mean tensors and between the two standard-deviation tensors are 0.9994 and 0.9973. A second prior extracted from 386 DropletVideo [58] clips has correlations of 0.9840 and 0.9816 with the InternVid mean and standard-deviation tensors. These measurements show similar frequency-wise statistics across the tested subsets and data sources.

Bounded amplitude correction. SpecLoR-Zero uses the interval $[\gamma\mu_{ref}, \mu_{ref} + k\sigma_{ref}]$ at each frequency bin. Amplitudes inside the interval remain unchanged; outliers receive the residual update in Eq. (3). Thus, the empirical statistics determine a correction range while retaining sample-specific amplitudes. This frequency-wise operation differs from the neighboring-bin averaging used by the smoothing control in Appendix §C.2.

Choosing Zero or Adapter. Zero provides a training-free rectifier after a one-time prior extraction and is convenient for deployment on another backbone using its VAE and target latent shape. Adapter predicts a residual from text, timestep, and lookahead context, using offline training with the base video model frozen. The two variants offer analytic and learned alternatives within the same sampling pipeline. Their architecture and training settings are given in Appendix §B.2.

A.3 Reconstruction at a Fixed Noise Endpoint

Let $x = z_{t,0}$ be the clean lookahead estimate, $\Delta = \hat{z}_{t,0} - z_{t,0}$ its rectification, and ϵ_{init} the initial noise. Applying the reconstruction in Eq. (5) to the unrectified and rectified estimates gives

$$\begin{aligned} \bar{z}_t &= (1-t)x + t\epsilon_{init}, \\ \hat{z}_t &= (1-t)(x+\Delta) + t\epsilon_{init}. \end{aligned} \tag{6}$$

The state change due to rectification within this reconstruction is therefore

$$\hat{z}_t - \bar{z}_t = (1-t)\Delta. \tag{7}$$

This update retains the timestep and selected noise endpoint. For a fixed seed, reusing ϵ_{init} makes the reconstruction deterministic. The lookahead and reconstruction control in Table 2 measures the effect of constructing $\bar{z}_t$, while the full method additionally applies Δ.

Direct inversion of Eq. (1) instead adds the previously evaluated velocity $v = v_\theta(z_t, t)$ to the rectified estimate. Define the local implied noise endpoint as $\epsilon_{loc} = x + v$. Since $z_t = x + tv$, direct inversion gives

$$\begin{aligned} \hat{z}_t^{inv} &= x + \Delta + tv = z_t + \Delta \\ &= (1-t)(x+\Delta) + t(\epsilon_{loc} + \Delta). \end{aligned} \tag{8}$$

The implied noise endpoint shifts with the sample-dependent correction. Reconstruction with Eq. (5) retains ϵ_{init} and applies the factor $(1-t)$ to the clean-anchor update.

A.4 Extended Discussion on Trajectory Control

Addressing trajectory drift in continuous-time generative models is a rapidly evolving area. Our proposed SpecLoR shares the foundational motivation of recent inference-time interventions, yet diverges significantly in its core mechanism.

Stochastic Spatial Perturbation vs. Deterministic Spectral Rectification. Recent advancements, most notably Self-Refining Video Sampling (SRVS) [13] and Restart Sampling [12], tackle trajectory drift by injecting stochastic noise into intermediate spatial latents. These methods effectively "rewind" the trajectory to re-evaluate structural features and dilute accumulated numerical errors. While SRVS intelligently minimizes accumulated error to guide this perturbation, the stochastic nature of noise injection within the highly entangled spatial domain can still inadvertently disrupt fragile local geometry. SpecLoR supports deterministic continuation with a fixed noise endpoint. A 3D FFT separates amplitude and phase, enabling an update to the empirical spectral energy distribution.

Spatial Optimization vs. Spectral Efficiency. Other concurrent works, such as FlowMo [47], enforce temporal coherence through variance-based guidance or gradient-based manipulation. While highly effective at maintaining motion consistency, spatial optimization often requires repeated forward passes or heavy gradient computations, substantially inflating the computational footprint (*e.g.*, total inference time of $1.7\times$ the baseline). SpecLoR fundamentally bypasses this bottleneck by leveraging the efficiency of spectral decoupling. Rectifying the amplitude spectrum via our lightweight SpecLoR-Zero introduces a mere 10% overhead (4 additional NFEs), shifting the trajectory correction paradigm from computationally heavy spatial guidance to highly efficient spectral anchoring.

B Implementation and Reproducibility

B.1 Data & Inference Pseudocode

Data Sourcing. Both the global prior extraction (SpecLoR-Zero) and the training of the context-aware adapter (SpecLoR-Adapter) utilize the exact same dataset subset, consisting of 2,000 randomly sampled high-quality video clips strictly sourced from the InternVid-10M-FLT dataset [59].

Code & Reproducibility. Our code is provided in the supplementary materials. To describe the core inference procedure, we provide a PyTorch-style pseudo-code for the SpecLoR inference step in Algorithm S1.

Algorithm S1 Pseudo-code for the SpecLoR Inference Step with deterministic continuation in a PyTorch-like style.

```
# Models & Variables:
# v_theta: Flow Matching velocity network (\eg, Wan2.2 DiT).
# freq_adapter: Optional trained DiT for context-aware residual amplitude.
# initial_noise: The exact initial Gaussian noise map sampled at t=1.
def speclor_step(z_t, t, text_emb, v_theta, freq_adapter, global_prior, initial_noise,
     lambda_weight=0.5):
    # 1. Lookahead Projection (Eq. 1)
    velocity = v_theta(z_t, t)
    z_t0_pred = z_t - t * velocity

    # 2. Spatiotemporal Frequency Decoupling via 3D FFT
    # z_t0_pred shape: [B, C, T, H, W]
    fft_z = torch.fft.fftn(z_t0_pred.float(), dim=(-3, -2, -1))
    amplitude = torch.abs(fft_z)
    phase = torch.angle(fft_z)

    # 3. Amplitude Rectification
    if freq_adapter is not None: # SpecLoR-Adapter (Option B)
        delta_A = freq_adapter(z_t0_pred, t, text_emb)
        A_rectified = (amplitude + delta_A).clamp(min=0)
    else: # SpecLoR-Zero Global Prior (Option A)
        # Note: mu_ref and sigma_ref are full 4D Tensors [C, T, H, W]
        mu_ref, sigma_ref = global_prior
        # Apply asymmetric trust region clipping (Eq. 2)
        gamma, k = 0.6, 3.0
        A_target = torch.clamp(amplitude, min=gamma * mu_ref, max=mu_ref + k * sigma_ref)
        # Apply bounded residual update (Eq. 3)
        A_rectified = amplitude + lambda_weight * (A_target - amplitude)

    # 4. Re-Noising & Inference
    # Reconstruct the spectrum keeping Phase strictly locked
    fft_rectified = A_rectified * torch.exp(1j * phase)
    z_t0_rectified = torch.fft.ifftn(fft_rectified, dim=(-3, -2, -1)).real

    # Anchor the trajectory back to the current timestep (Eq. 5)
    z_t_rectified = (1 - t) * z_t0_rectified + t * initial_noise

    return z_t_rectified
```

B.2 Extended Implementation Details

B.2.1 Global Prior Statistics (SpecLoR-Zero)

The 2,000 sampled videos from InternVid-10M-FLT are preprocessed to a spatial resolution of 480×832 with a sequence length of 81 frames. We encode these clean videos using the frozen Variational Autoencoder (VAE) of Wan2.2-A14B [2]. To ensure numerical precision during frequency decoupling, latent representations are explicitly cast to `float32` before the 3D FFT. Crucially, rather than averaging the spectrum over spatiotemporal dimensions to obtain a single scalar per channel, we preserve the frequency-wise statistics. Thus, the extracted global prior mean μ_{ref} and standard deviation σ_{ref} are highly detailed 4D tensors of shape $C \times T \times H \times W$ (specifically, $16 \times 21 \times 60 \times 104$).

B.2.2 SpecLoR-Adapter Architecture and Training

To capture context-aware dynamic variance, SpecLoR-Adapter is designed as a highly lightweight architecture constructed from merely 2 Diffusion Transformer (DiT) [25] blocks. Instead of processing spatial latents directly, the adapter performs a 3D FFT on the lookahead latent $z_{t,0}$. The resulting magnitude spectrum undergoes a logarithmic mapping, $\log(1 + \mathcal{A})$, before projection into a compact hidden dimension of 128. Rotary Positional Embeddings (RoPE) are explicitly applied along the frequency dimension. Cross-attention layers within each 4-head DiT block condition the residual amplitude prediction on text embeddings (context dimension of 4096). We employ the AdamW optimizer (learning rate 5×10^{-5}, cosine scheduler, 300 warmup steps) to minimize the MSE between predicted and ground-truth logarithmic magnitudes. The model is trained for 30 epochs (batch size 1, 4 gradient accumulation steps) in `bfloat16` precision, while the foundational Wan2.2

parameters remain strictly frozen. The corrected amplitude is clamped to be nonnegative before the log-amplitude loss and inverse FFT.

B.2.3 Memory Overhead of Spatiotemporal FFT

While SpecLoR achieves exceptional computational efficiency regarding NFEs, computing the full 3D FFT via `torch.fft.fftn` introduces a memory footprint during inference. For a 14B model like Wan2.2, standard latents ($z_{t,0}$) are in 16-bit float (2 bytes/element). The implementation casts these real tensors to `float32` before the FFT, producing `complex64` coefficients (32-bit real + 32-bit imaginary, totaling 8 bytes/element). Given a batch size B, channels C, and spatiotemporal dimensions $T \times H \times W$, the storage required for the complex FFT output is:

$$M_{FFT} = B \times C \times T \times H \times W \times 8 \text{ bytes}. \tag{9}$$

While perfectly manageable for standard resolutions (*e.g.*, 480×832, 81 frames), this linear scaling with sequence length and spatial pixel count highlights a potential bottleneck for future ultra-high-resolution or minutes-long video generation, necessitating future explorations into windowed or patch-based spectral approximations.

B.3 Configuration of Additional Evaluations

The additional evaluations reuse the prior extraction and inference components described in Appendices §B.1 and §B.2.

The core inference procedure is given in Algorithm S1.

Unless specified otherwise, the additional Wan2.2 experiments use VideoJAM-Bench, with SpecLoR-Zero configured as in Sec. 4.3 ($\gamma = 0.6$, $k = 3$, and $\lambda = 0.5$).

B.4 Licenses of Assets Used

To ensure full compliance with the NeurIPS guidelines regarding the use of existing assets, we document the licenses and terms of use for all pre-trained models, datasets, and evaluation benchmarks utilized in this research:

- **Wan2.2 Model Series:** The foundational text-to-video model, Wan2.2-A14B [2], is distributed under the **Apache 2.0 License**, which permits free use, modification, and distribution for both research and commercial purposes.
- **InternVid-10M-FLT Dataset:** The 2,000 video clips used for extracting the global spectral prior and training the SpecLoR-Adapter were sampled from the InternVid dataset [59], which is available under the **CC BY-NC-SA 4.0 License**.
- **Evaluation Benchmarks & Metrics:** The standardized video evaluation frameworks and metric models employed in our study—including VBench [54], VideoScore2 [53], Dynamic-Bench [13], VideoJAM-Bench [52], and MovieGen-Bench [60]—are generally open-sourced under the **Apache 2.0 License**, **CC BY-NC 4.0 License** or **MIT License** for research purposes.

We confirm that our usage of these assets is strictly limited to research and evaluation, fully respecting the original creators' terms of service and copyright notices.

C Controlled Experiments and Ablations

C.1 Validation against Real-Video References

C.1.1 Real-Video Reference Distributions

We use 20 DropletVideo [58] video-caption pairs, disjoint from the InternVid prior-estimation set, with three seeds per video. We encode each real video and add the scheduled noise to construct reference states. Generated Wan2.2 states use the corresponding captions and matched noise. We evaluate the resulting distributions with squared maximum mean discrepancy (MMD^2). Averaged over Steps 2–5, the values are 0.0192 for the current latent, 0.0669 for the clean lookahead, and 0.3616 for its amplitude; all reported tests give $p < 0.001$. These measurements quantify distribution differences in the respective representations.

C.1.2 Correction against a Known Clean Reference

In a separate correction test, we apply Eq. (1) to each noised real-video state and rectify the predicted lookahead amplitude with SpecLoR-Zero. The known clean latent supplies the evaluation reference. Mean absolute log-amplitude error decreases from 1.237 to 0.832, a reduction of approximately 33%, with improvement on all 20 videos. This directly measures the effect of rectification on amplitude prediction error.

C.2 Amplitude Smoothing Control

The smoothing control applies phase-preserving average pooling to the amplitude with the same intervention window and update strength as SpecLoR-Zero. Table S1 tests whether generic smoothing explains the gains. Smoothing achieves the highest Motion Smoothness, while both SpecLoR variants produce higher Visual Quality and Physical Consistency on both benchmarks.

Table S1: Amplitude rectification versus amplitude smoothing. The smoothing control uses the same intervention window and update strength as SpecLoR-Zero.

Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑
			VideoJAM-Bench		
Amplitude smoothing	3.984	3.310	3.825	**0.941**	**0.988**
SpecLoR-Zero	4.113	3.432	3.909	0.937	0.974
SpecLoR-Adapter	**4.123**	**3.441**	**3.924**	**0.941**	0.976
			Dynamic-Bench		
Amplitude smoothing	3.597	2.835	3.773	**0.918**	**0.987**
SpecLoR-Zero	**3.651**	2.993	3.810	0.915	0.980
SpecLoR-Adapter	3.585	**3.056**	**3.861**	0.914	0.981

C.3 Frequency Correspondence of the Prior

The shuffled-prior control retains the prior values but assigns them to incorrect frequency locations. Table S2 isolates the importance of this correspondence on VideoJAM-Bench. Across all three VideoScore2 metrics, SpecLoR-Zero scores above the baseline, while shuffling the prior scores below it. The shuffled prior also introduces local structural artifacts, including malformed or additional limbs. These results support the use of frequency-specific amplitude statistics.

Table S2: Effect of prior-frequency correspondence on VideoJAM-Bench. Shuffling changes the frequency assignments while retaining the same prior values.

Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑
SpecLoR-Zero	**4.113**	**3.432**	**3.909**
Baseline	4.062	3.237	3.873
Shuffled prior	3.965	3.224	3.788

C.4 Sensitivity to the Lookahead Window

In the main text, we rigidly restrict the SpecLoR intervention to steps 2 to 5 of a 40-step UniPC schedule. Table 3 empirically demonstrates the benefit of isolating this specific early, high-noise regime.

The Cost of Late Intervention. Delaying the rectification window to later stages (*e.g.*, steps 10-13 or 14-17) reduces its effectiveness. As shown, Consistency progressively deteriorates to 0.933 and 0.932, while Motion Smoothness falls to 0.971 and 0.970, performing demonstrably worse than the unguided baseline. These results favor amplitude rectification during the early formation of global motion structure.

The Early Advantage. Conversely, anchoring the amplitude specifically during steps 2-5 successfully guides the trajectory before severe structural drift solidifies, achieving highly competitive Consistency

(0.939) and Motion Smoothness (0.973) with the training-free prior, and peaking at 0.942 and 0.978 using our SpecLoR-Adapter.

D Extended Evaluation

D.1 Additional Baseline Comparisons

D.1.1 SRVS at a Matched 44-NFE Budget

We evaluate SRVS [13] at 44 NFEs using the same seed as both SpecLoR variants (Table S3). At this budget, SpecLoR-Adapter scores higher on all five VideoJAM metrics, and SpecLoR-Zero scores higher on four. Both variants score higher than SRVS on all five Dynamic-Bench metrics. The 60-NFE SRVS and NFE$\times$1.5 results in Table 1 provide higher-budget comparisons. FlowMo [47] uses its official test-time optimization setting, with the reported $1.7\times$ inference time.

Table S3: SRVS and SpecLoR evaluated with the same seed at 44 NFEs.

Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑
			VideoJAM-Bench		
SRVS (44 NFEs)	4.077	3.371	3.881	0.940	0.973
SpecLoR-Zero	4.113	3.432	3.909	0.937	0.974
SpecLoR-Adapter	**4.123**	**3.441**	**3.924**	**0.941**	**0.976**
			Dynamic-Bench		
SRVS (44 NFEs)	3.557	2.837	3.760	0.900	0.979
SpecLoR-Zero	**3.651**	2.993	3.810	**0.915**	0.980
SpecLoR-Adapter	3.585	**3.056**	**3.861**	0.914	**0.981**

D.1.2 Frequency Initialization and Guidance Methods

We adapt FreeInit [11] and FreqPrior [17] from their published methods and available diffusion-backbone implementations to the Wan2.2 sampler, using 80 and 98 NFEs, respectively. We also evaluate CFG-Zero* [43] and Enhance-A-Video [32] under the same Wan2.2 setup. Table S4 reports VideoJAM-Bench results. At 44 NFEs, both SpecLoR variants achieve higher T2V Alignment and Physical Consistency than these methods. FreeInit obtains higher Visual Quality, and the comparison methods obtain higher Motion Smoothness. These results illustrate different balances between physical structure, appearance, and temporal metrics.

Table S4: Additional inference-time methods on VideoJAM-Bench with Wan2.2. FreeInit and FreqPrior use 80 and 98 NFEs; both SpecLoR variants use 44 NFEs.

Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑
FreeInit	4.024	3.355	**3.936**	0.948	0.986
FreqPrior	3.911	3.169	3.604	**0.951**	0.981
CFG-Zero*	3.988	3.371	3.826	0.942	**0.990**
Enhance-A-Video	4.070	3.418	3.879	0.943	0.983
SpecLoR-Zero	4.113	3.432	3.909	0.937	0.974
SpecLoR-Adapter	**4.123**	**3.441**	3.924	0.941	0.976

D.2 Motion Coherence on MovieGen-Bench

Beyond structural physics in local interactions, we further investigated whether amplitude rectification improves overall motion stability in high-fidelity, high-resolution cinematic generations. As presented in Table S5, both SpecLoR-Zero and SpecLoR-Adapter exhibit substantial gains on the rigorous MovieGen-Bench [60], comprehensively lifting the Dynamic Degree and Motion Smoothness without sacrificing base Consistency. These results demonstrate smoother motion under spectral amplitude rectification. The context-aware SpecLoR-Adapter achieves the highest Dynamic Degree while maintaining base Consistency.

Table S5: Quantitative results on MovieGen-Bench. Both variants of SpecLoR substantially lift the Dynamic Degree and Motion Smoothness with minimal computational overhead, supporting transfer to cinematic video generation.

Method	Consistency↑	Motion Smooth.↑	Dynamic Degree↑	NFE↓	Time↓
Wan2.2 T2V	0.932	0.977	0.754	40	
+ NFE×1.5	0.932	0.977	0.757	60	1.5×
+ SpecLoR-Zero (Ours)	**0.933**	**0.978**	0.782	44	1.1×
+ SpecLoR-Adapter (Ours)	0.932	**0.978**	**0.783**	44	1.1×

Table S6: Quantitative comparison on `HunyuanVideo-1.5`. Best scores are in **bold**.

	VideoScore2			VBench		Efficiency	
Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑	NFE↓	Time↓
Results on VideoJAM-Bench [52]							
HunyuanVideo-1.5	3.833	2.924	3.486	0.938	0.985	50	
+ SpecLoR-Zero (Ours)	**3.890**$_{\pm 0.015}$	**3.260**$_{\pm 0.017}$	**3.736**$_{\pm 0.012}$	**0.943**$_{\pm 0.003}$	**0.992**$_{\pm 0.002}$	57	1.14×
Results on Dynamic-Bench [13]							
HunyuanVideo-1.5	3.213	2.634	3.532	0.911	0.987	50	
+ SpecLoR-Zero (Ours)	**3.283**$_{\pm 0.011}$	**2.785**$_{\pm 0.018}$	**3.674**$_{\pm 0.014}$	**0.915**$_{\pm 0.004}$	**0.993**$_{\pm 0.001}$	57	1.14×

D.3 Generalization across Video Backbones

D.3.1 HunyuanVideo-1.5

To verify the model-agnostic nature of SpecLoR, we extend the framework to the `HunyuanVideo-1.5` (480p) architecture [26]. Due to time and computational constraints, we focus our evaluation in this section exclusively on the training-free `SpecLoR-Zero` variant.

Furthermore, during inference, we empirically adjusted the default Classifier-Free Guidance (CFG) scale down to 3.0. While the baseline `HunyuanVideo` model often operates at higher guidance scales (e.g., 6.0), applying our strict spectral amplitude rectification under such aggressive CFG introduces severe visual artifacts. This occurs because an overly high CFG actively pushes latents into saturated, high-variance regions, creating a mathematical conflict with `SpecLoR-Zero`'s objective of anchoring macroscopic energy back to a natural prior. By lowering the CFG to 3.0, we effectively harmonize the text guidance with our spectral regularization, yielding coherent and high-fidelity dynamics.

For the 50-step sampling schedule used in HunyuanVideo-1.5, we set the Lookahead Window to steps 4–10. This configuration is chosen for two primary reasons:

- **Efficiency Consistency:** Applying 7 additional steps in a 50-step schedule incurs a modest 14% computational overhead, keeping the relative cost comparable to our 40-step Wan2.2 experiments.
- **Timestep Alignment:** In the discrete flow-matching process, steps 4–10 in a 50-step schedule correspond to a noise-level range (timesteps) that is more equivalent to steps 2–5 in a 40-step schedule. This ensures that the spectral rectification targets the same critical early stage where global motion structures begin to coalesce.

Preliminary results (See Table S6) confirm that SpecLoR successfully suppresses structural artifacts and improves motion stability in HunyuanVideo-1.5 without modifying the backbone architecture.

D.3.2 LTX-Video

We evaluate SpecLoR-Zero on `LTX-Video-0.9.7-dev` [61], extending the backbone evaluation to a compressed latent representation and a coarse-to-fine generation pipeline. The first pass uses 30 sampling steps at 512×768. The pipeline then upscales and refines the video at 1024×1536 with four effective denoising steps, followed by resizing to 768×1152. We apply SpecLoR-Zero only at Steps 2–4 of the first pass and extract the spectral prior using the LTX VAE.

Table S7 reports all five metrics on VideoJAM-Bench. Visual Quality increases from 3.668 to 3.866 and Physical Consistency from 2.992 to 3.349. Consistency increases from 0.959 to 0.960, while T2V Alignment changes from 3.857 to 3.799 and Motion Smoothness from 0.994 to 0.993. This

evaluation demonstrates transfer of the Zero rectification pipeline to LTX with a model-specific prior and intervention window.

Table S7: LTX-Video-0.9.7-dev results on VideoJAM-Bench. SpecLoR-Zero is applied at Steps 2–4 of the first generation pass.

Method	T2V Align.↑	Phys. Cons.↑	Visual Qual.↑	Consistency↑	Motion Smooth.↑
LTX-Video	**3.857**	2.992	3.668	0.959	**0.994**
+ SpecLoR-Zero	3.799	**3.349**	**3.866**	**0.960**	0.993

D.4 Human Preference Study

D.4.1 Study Design and Evaluation Protocol

Five graduate volunteers with computer vision backgrounds independently evaluate the same 128 randomized, anonymized video pairs on VideoJAM-Bench, comprising 256 videos. Each pair compares the Wan2.2 baseline with SpecLoR-Adapter using Stochastic Contraction ($\epsilon \sim \mathcal{N}(0, I)$, with fresh noise sampled at each re-noising step) and is judged for Text Alignment, Motion Coherence, and Video Quality. This gives 640 judgments per criterion and 1,920 judgments in total.

To ensure a strict double-blind setup, we employed a local evaluation protocol. A custom Python script randomly shuffled and renamed the video pairs (Baseline vs. SpecLoR) to eliminate positional bias. Evaluators viewed these anonymized pairs locally and recorded their preferences (Video 1, Video 2, or Tie) across Text Alignment, Motion Coherence, and Video Quality, as illustrated in Fig. S2. The randomization script is provided in the supplementary materials.

	A	B	C	D	E
	Folder_Name	Text Alignment(1/2/Tie)	Motion Coherence(1/2/Tie)	Video Quality(1/2/Tie)	Prompt
	000_A_woman_performing_an_intricate				A woman

Figure S2: An example of the local scoring spreadsheet used by evaluators to record their preferences across three criteria for each anonymized video pair.

D.4.2 Preference across Evaluators

Table S8 reports the Ours/Tie/Baseline counts for each evaluator. All five evaluators give more votes to SpecLoR than to the baseline on each criterion.

Table S8: Individual evaluator votes, shown as Ours / Tie / Baseline. Each evaluator assesses 128 pairs per criterion.

Evaluator	Text Alignment	Motion Coherence	Video Quality
A	23 / 93 / 12	47 / 64 / 17	57 / 46 / 25
B	24 / 92 / 12	53 / 51 / 24	53 / 54 / 21
C	29 / 79 / 20	50 / 57 / 21	55 / 51 / 22
D	27 / 87 / 14	49 / 61 / 18	52 / 56 / 20
E	27 / 89 / 12	56 / 47 / 25	58 / 43 / 27

D.4.3 Agreement and Prompt-Level Statistical Analysis

Agreement is computed over all three labels using mean pairwise agreement and Fleiss' κ. For significance testing, each prompt is one comparison unit, with its five judgments kept together. We use an exact two-sided prompt-level permutation test with Holm correction across the three criteria. The 95% confidence intervals are obtained by resampling prompts. In Table S9, preference is the share of SpecLoR votes among non-tie judgments. Ties contribute to the vote counts and agreement analysis and have zero directional contribution to the preference test.

The preference for SpecLoR is significant for all three criteria after correction, with adjusted $p = 0.038$ for Text Alignment and $p < 0.001$ for Motion Coherence and Video Quality. Figure 4 reports percentages among all judgments; its white error bars show standard deviation across evaluators.

Table S9: Human preference and agreement on VideoJAM-Bench. Preference and its confidence interval use non-tie judgments; O/T/B gives the complete Ours/Tie/Baseline counts.

Criterion	Agreement	κ	O / T / B	Preference (95% CI)	Adjusted p
Text Alignment	85.3%	0.690	130 / 440 / 70	65.0% [51.2%, 77.6%]	0.038
Motion Coherence	81.9%	0.709	255 / 280 / 105	70.8% [60.9%, 80.1%]	< 0.001
Video Quality	84.4%	0.752	275 / 250 / 115	70.5% [60.8%, 79.7%]	< 0.001

D.5 Failure Cases and Limitations

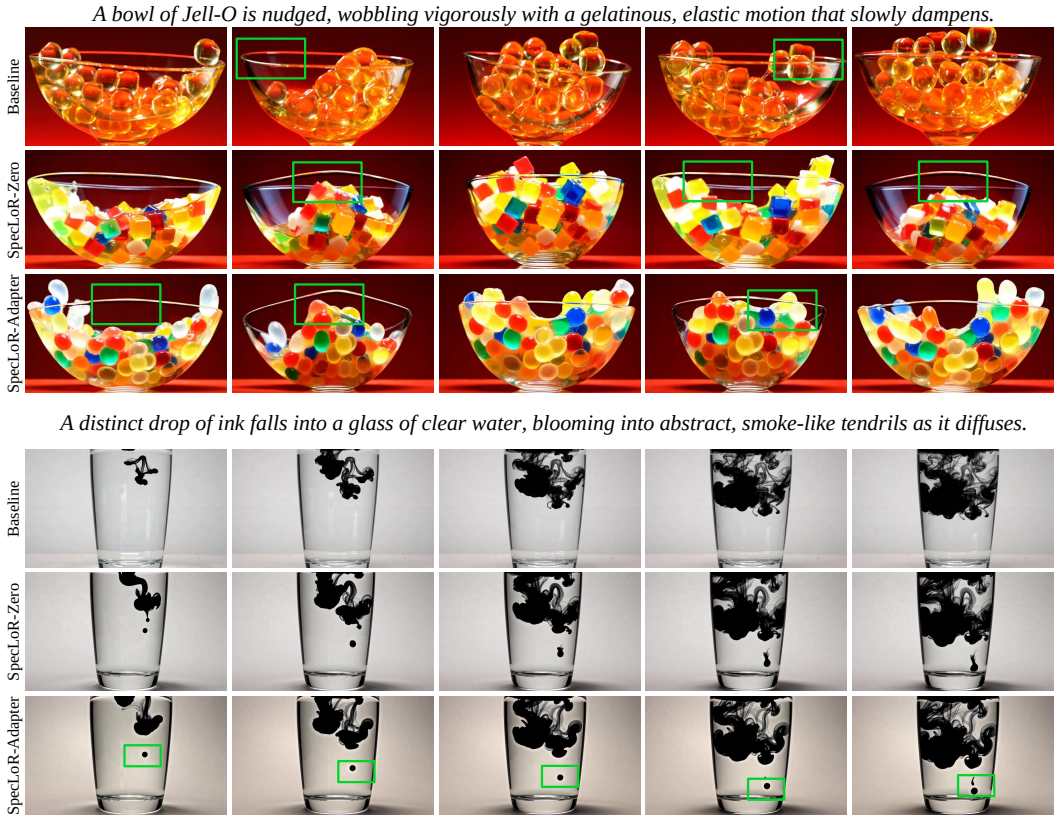

Figure S3: Visual demonstration of failure cases under chaotic, non-rigid transformations. The SpecLoR outputs exhibit deformation of a rigid glass bowl during gelatinous motion (top) and a rigid ink-drop shape during fluid diffusion (bottom).

While SpecLoR excels at preserving kinematic structure for rigid bodies and human dynamics, it exhibits failure modes in highly chaotic systems. Fig. S3 visually dissects these limitations under extreme non-rigid transformations.

Elastic Dynamics (Jell-O). When evaluating gelatinous, wobbly motion, the generated motion can extend elastic deformation into the surrounding environment. As shown in the top row of Fig. S3, the rigid glass bowl itself begins to drastically warp and wobble in the SpecLoR output, adopting the elastic properties of the Jell-O.

Fluid Diffusion (Ink). In scenarios involving topological changes, such as an ink drop diffusing in water, the generated ink drop can retain an unnatural rigid structure. Rather than blooming into smoke-like tendrils, the ink droplet persists as an unnatural, rigid spherical structure falling through the water (bottom row, green boxes), failing to capture the continuous spreading of the ink.

We empirically observe that employing the stochastic variant of our method partially mitigates this limitation. This suggests that future trajectory control mechanisms must distinguish between rigid structural preservation and non-rigid chaotic evolution.

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 our algorithmic contributions regarding the SpecLoR framework and explicitly scope its improvements to motion coherence and computational efficiency.

   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: We explicitly address the limitations of our method, including potential phase-locking artifacts and 3D FFT memory overhead, in the newly added Discussion and Conclusion section.

   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: This paper focuses on an empirical, training-free algorithmic contribution for inference-time sampling and does not introduce formal theoretical 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: Complete implementation details, including the base model, ODE solver configurations, and specific lookahead window steps, are clearly specified in Section 5.1.

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: Our code is provided in the supplementary materials. Data sources and implementation details are described in Appendix §B.

   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: All key hyperparameters, such as intervention steps, modulation strength, and benchmark evaluation protocols, are detailed in the experimental setup and subsequent ablation sections.

   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: [Yes]

   Justification: We report quantitative performance averaged across comprehensive, standardized benchmarks (e.g., Dynamic-Bench, VideoJAM-Bench) to ensure robust and statistically meaningful evaluation.

   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: Section 5.1 explicitly states the hardware utilized (910B NPU with 60GB memory) and the approximate inference time required per video generation.

   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 research conducted fully adheres to the NeurIPS Code of Ethics, focusing responsibly on correcting trajectory drift in existing generative models.

   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: Section 6 discusses dual-use risks, including deepfakes and misinformation, and outlines mitigation strategies for responsible deployment.

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: [Yes]

    Justification: Section 6 recommends the concurrent development of resilient latent watermarking and video forgery detection to mitigate potential societal harms.

    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: All utilized pre-trained models (*e.g.*, Wan2.2), benchmark datasets, and evaluation metrics are properly cited with appropriate respect to their original licenses.

    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: This work introduces an inference-time algorithmic framework and does not release new datasets or novel pre-trained foundation models.

    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: [Yes]

    Justification: The user study involved basic visual preference annotations (A/B testing) for generation quality.

    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: [No]

    Justification: The visual preference A/B testing conducted for evaluating video quality does not qualify as sensitive human subjects research requiring IRB approval under standard institutional guidelines.

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: Large Language Models were not utilized as an original or non-standard component of the core generative methodology in this research.

    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.