Physics in 2-Steps: Locking Motion Priors Before Visual Refinement Erases Them

Anonymous Authors[1]

Abstract

Image-to-Video diffusion models leverage input images to generate visually stunning content, yet frequently produce motion that violates physical laws. We reveal a surprising finding: a 2-step generation often exhibits *better* physical consistency than a 50-step output from the same model. Through spectral analysis, we trace this to phase erosion during denoising; the phase degrades significantly (dropping by $\approx 18\%$ from step 2 to step 50), whereas the magnitude remains relatively stable. Building on this insight, we propose **PhaseLock**, a training-free framework that locks the valid motion priors into the denoising trajectory found in few-step inference. Rather than requiring 50 steps to establish physics, PhaseLock extracts a motion prior from just 2 steps and enforces it onto high-fidelity generation via *Latent Delta Guidance*. Our approach effectively prevents phase degradation, achieving both high visual fidelity and superior physical consistency scores. Extensive experiments demonstrate an average improvement of 6.2 points across diverse models with negligible overhead ($1.06\times$ time, $1.02\times$ memory), eliminating the need for expensive external guidance methods ($\sim 5\times$ time).

1. Introduction

Recent advances in video generation have achieved remarkable progress in producing visually coherent and semantically aligned content (Blattmann et al., 2023a; Ho et al., 2022; Brooks et al., 2024). Modern diffusion-based models demonstrate strong capabilities in understanding *what* should appear in a video; objects, scenes, and their visual attributes are rendered with impressive fidelity. Yet, despite this semantic competence, a critical limitation persists: *physical hallucination*, the generation of motion that violates fundamental physical laws (Kang et al., 2024; Bansal et al., 2024; Chow et al., 2025). While substantial efforts have been devoted to embedding external physical knowledge through physics engines or external modules (Liu et al., 2024; Yuan et al., 2023; Zhang et al., 2025b; Yuan et al., 2026) or scaling datasets (Assran et al., 2025; Yang et al., 2024; Wang et al., 2025), these approaches often demand excessive computational overhead or human annotation. Despite their scale, they continue to produce physically implausible dynamics, impeding the path toward reliable world simulators (Qin et al., 2024). This raises a key question: *does the model lack physical knowledge, or does it forget what it already knows?*

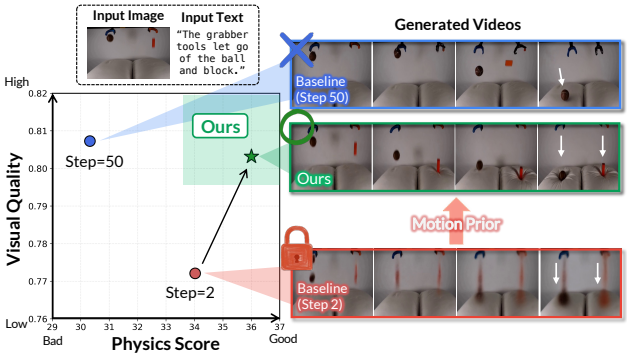


Figure 1. **Overview of PhaseLock.** Few inference ($T = 2$) captures accurate physical motion (following the white arrow) but lacks textural detail, whereas standard inference ($T = 50$) achieves photorealism but compromises physical integrity with hallucinations. Our method, PhaseLock, extracts the valid motion prior from the few inference stage and inject it during the denoising steps, ensuring both high visual fidelity and physical consistency. Sec. 3.1 details the specific experimental setup.

To investigate the root cause of physical failure, we focus on Image-to-Video (I2V) generation. While text prompts can successfully induce correct dynamics when the motion is explicitly described (*e.g.,* "a ball falling"), they frequently fail when the prompt is ambiguous, describing a causal setup (*e.g.,* "letting go of a ball") without specifying the resulting motion. To address this ambiguity, we turn to the input image, which offers more grounded physical cues than high-level text. Drawing on findings that visual observation sup-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ports intuitive physics (Piloto et al., 2022; Gao et al., 2025), prior work shows that reference images implicitly encode material properties, structural constraints, and instantaneous motion (Li et al., 2023). Despite this rich availability of visual priors, however, we observe that current models still suffer from severe physical hallucinations.

We hypothesize that physical priors encoded in the image fail to propagate due to structural loss during denoising. Building on insights from diffusion dynamics literature (Song et al., 2020; Qi et al., 2023), which demonstrate that early steps focus on global motion and structure rather than high-frequency details, we conducted a comparative analysis between few and standard inference steps. We begin with a counter-intuitive observation: a video generated with extremely few inference steps (*e.g.,* only 2 steps) often exhibits better physical consistency than a standard 50-step generation, as shown in Fig. 1. While the standard 50-step output yields high visual fidelity, it often hallucinates erratic motions or vanishing objects (*e.g.,* failing to capture the correct vertical drop shown in Fig. 1). Conversely, the few step inference preserves the physical trajectory, a finding validated by improvements in physics scores across Physics-IQ Benchmark (Motamed et al., 2025). Considering that both outputs are generated from the identical model, seed, and conditioning, the fact that the motion trajectory diverges so drastically solely due to the increased step count reveals a noteworthy trade-off: *the model correctly retrieves the valid "motion prior" in the few step generation, but inadvertently overwrites this physical structure during the pursuit of visual refinement.*

To understand the origin of this degradation, we analyze the generation process in the frequency domain. Decomposing the video latent into magnitude and phase components reveals that while the magnitude spectrum (encoding texture and contrast) remains stable, the phase spectrum (encoding structural dynamics) degrades significantly (dropping by $\approx$ 18% from step 2 to step 50). A causal analysis further confirms that motion dynamics are inherently phase-sensitive; for instance, a 50% phase corruption induces $8.5\times$ greater optical flow distortion than an equivalent magnitude corruption. Therefore, locking the phase at small timesteps enables us to prevent physics hallucination without training, by effectively leveraging the inherent motion priors.

Based on these findings, we propose **PhaseLock**. Grounded in the observation that physical consistency is established within just 2 steps, our *Latent Delta Guidance* leverages the few inference latent as a motion prior. Specifically, we employ a straightforward mechanism that computes inter-frame deltas from the 2-step latent and applies these differences to guide the denoising process. This approach effectively preserves the phase information throughout the high-fidelity denoising process. Our method demonstrates an average improvement of 6.2% across diverse models. While WM-Reward (Yuan et al., 2026) shows comparable performance gains, our approach achieves this with much lower computational cost ($1.06\times$ time, $1.02\times$ memory), whereas WM-Reward requires substantially more resources ($\sim 5\times$ time).

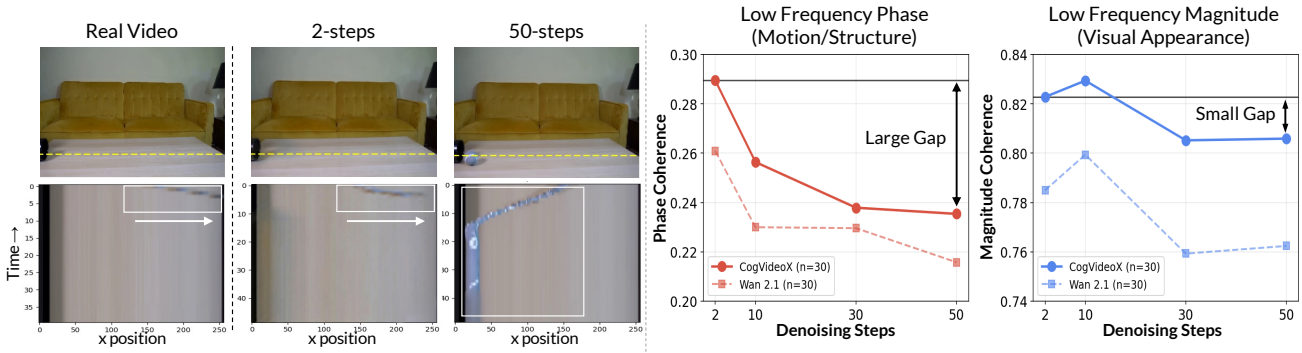
Our contributions are summarized as follows:

- We find that few inference timestep yield better physical structure, revealing that the denoising process progressively erodes the phase spectrum (encoding structural dynamics).
- We propose PhaseLock, a model-agnostic, training-free strategy that locks the physical priors captured in few inference steps (NFE = 2), effectively preserving the structural phase spectrum.
- Our method achieves strong physical consistency (avg. +6.2%) with marginal overhead ($1.06\times$ time, $1.02\times$ memory), eliminating the need for expensive external guidance ($\sim 5\times$ time).

2. Related Works

Physical Consistency in Video Generation. Large-scale video diffusion models (Ho et al., 2022; Blattmann et al., 2023b; Singer et al., 2023; Bar-Tal et al., 2024; Peebles & Xie, 2023) have achieved remarkable visual quality, yet consistently struggle with physical consistency (Li et al., 2025; Meng et al., 2025). To more accurately assess physical plausibility, benchmarks such as VideoPhy (Bansal et al., 2024) and PhyGenBench (Meng et al., 2025) have employed VLM-based evaluation protocols. To rigorously evaluate physical understanding, Physics-IQ (Motamed et al., 2025) was introduced as a specialized benchmark that distinguishes genuine physical dynamics from mere visual realism. To improve physical consistency, several distinct directions have emerged (Xue et al., 2025; Zhang et al., 2025a). WISA (Wang et al., 2025) curates physics-specific datasets, yet this process is prohibitively expensive and often lacks generalization (Kang et al., 2024). Alternatively, methods like PhysGen (Liu et al., 2024) and VideoREPA (Zhang et al., 2025b) incorporate external simulators or model alignment More recently, WMReward (Yuan et al., 2026) uses a latent world model as a physics reward for test-time trajectory optimization, but suffers from prohibitive overhead.

Frequency Analysis and Diffusion Dynamics. Signal processing perspectives have provided key insights into diffusion behavior. Foundational works like EDM (Karras et al., 2022) and Cold Diffusion (Bansal et al., 2023) established coarse-to-fine generation dynamics, where global structures form early and high-frequency details emerge later. This perspective has been formalized as spectral autoregression (Dieleman, 2024), revealing diffusion models



(a) Space-Time Slices Showing Motion Trajectories

(b) Low Frequency Spectral Analysis

Figure 2. **Analysis of physical degradation across denoising steps.** We compare the baseline models (CogVideoX, WAN2.1) at few ($T = 2$) and default ($T = 50$) inference steps against the real video. *(a) Spatio-temporal ($x - t$) slices:* The yellow line indicates the temporal reference axis. As highlighted in the white box, both the real video and Step 2 accurately follow the physical trajectory. In contrast, Step 50 exhibits severe physical hallucination, moving in the opposite direction. *(b) Spectral evolution analysis:* We analyze the low-frequency components of Phase and Magnitude across steps $t \in \{2, 10, 30, 50\}$. While magnitude remains consistent between Step 2 and Step 50, phase significantly diverges, identifying phase corruption as the source of the motion artifacts.

implicitly operate in the frequency domain. Specifically, FreeU (Si et al., 2024) reweights backbone and skip connections by frequency, FreeInit (Wu et al., 2024) refines low-frequency initialization for temporal consistency, and FreqPrior (Yuan et al., 2025) filters noise in the frequency domain. Moving beyond frequency-band analysis, we identify 'phase erosion' as the mechanism behind hallucinations and lock phase dynamics to ensure physical consistency.

3. Mechanism Analysis

Here, we systematically analyze the mechanism of physical hallucination through a step-wise examination of video generation. We observe a trade-off between visual refinement and dynamic consistency, and identify the loss of phase information as a primary cause. Further details on the analysis setup are provided in §Appendix B.1.

3.1. Divergence of Visual and Physical Fidelity

Motivated by the inherent coarse-to-fine generation dynamics of diffusion models (Ho et al., 2022; Choi et al., 2022), where global structure is established before high-frequency details, we hypothesize that physical motion priors are captured in the few denoising process. To test this, we compare the generation results at the few inference steps ($T = 2$) versus the fully denoised steps ($T = 50$). As illustrated in Fig. 1, we observe a noticeable divergence between visual and physical fidelity. Specifically, comparing step 2 and step 50, the visual quality relative to real videos improves (LPIPS↓ (Zhang et al., 2018): $0.23 \rightarrow 0.19$), yet the physical integrity significantly degrades (Physics-IQ ↑: $34.02 \rightarrow 30.32$). This quantitative trend aligns with our qualitative observations as shown in Fig. 1. To scrutinize the motion dynamics, we examined the spatiotemporal (x-t) slices shown in Fig. 2(a). This visualization reveals that the 2-step result exhibits motion patterns most similar to the real video. In contrast, the 50-step video suffers from temporal inconsistencies, such as ball moving backward. Additional visualizations are provided in the §Appendix B.1.

This observation offers a crucial insight: the model inherently possesses implicit physical knowledge, which manifests as accurate coarse-level dynamics in the few inference steps. However, this structural information is eroded as the model focuses on high-frequency textural refinement. Although extending the denoising process beyond 50 steps may seem beneficial, we found that it improves physical consistency about 1% despite the substantial increase in inference time. See §Appendix D.2 for details.

3.2. Spectral Mechanism

To identify the mechanism behind the physical information loss, we analyze the generation process in the frequency domain. We decompose a video latent z into its magnitude and phase components via Fourier Transform: $\mathcal{F}(z) = A \cdot e^{i\phi}$. Following signal processing theory (Oppenheim et al., 1999), the phase spectrum ϕ preserves the spatio-temporal semantics (structural layout and motion trajectories), whereas the magnitude spectrum A captures the low-level statistics (texture, contrast).

Spectral Decomposition. We analyze the spectral dynamics of CogVideoX (Yang et al., 2024) and Wan 2.1 (Wan et al., 2025), specifically focusing on the low-frequency region (normalized distance < 0.4) of the 3D spatio-temporal spectrum. Given that high-frequency details are scarce in Step 2, the low-frequency region serves as the primary source for motion dynamics. To quantify the fidelity of the

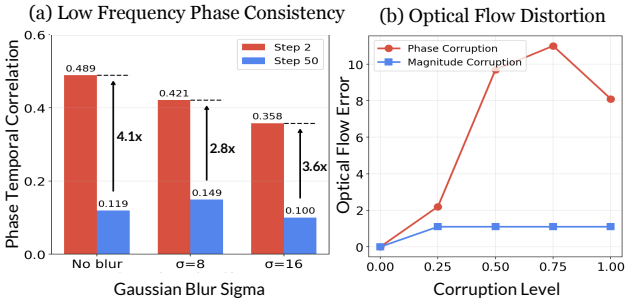


Figure 3. **Further analysis on phase properties.** *(a) Blur control:* Even with Gaussian blur applied to match sharpness, Step 2 retains significantly higher Phase Temporal Correlation, proving that phase loss is structural rather than a frequency artifact. *(b) Phase Sensitivity:* Physical dynamics are highly sensitive to Phase corruption, degrading rapidly compared to the stable Magnitude.

generated content, we employ two metrics: (1) Phase Coherence, defined as the mean cosine similarity between the phase angles of the real ground truth video's latents and generated video's latents, and (2) Magnitude Correlation, measured via the Pearson correlation of their log-magnitudes. As illustrated in Fig. 2 (b), while the Magnitude Correlation remains relatively stable with minimal loss (decreasing only by ∼2-3%), the Phase Coherence exhibits a sharp degradation, dropping by approximately 18% in both CogVideoX and Wan 2.1 models. We observe that motion dynamics are closely related to the phase spectrum, suggesting that visual refinement compromises this essential information. The Spectral analysis results with our method applied are provided in §Appendix D.4.

Low-Frequency Structural Integrity. It could be argued that the high phase consistency of Step 2 is merely an artifact of its inherent blurriness, which lacks high-frequency disturbances. To rule out this possibility, we applied varying degrees of Gaussian blur ($\sigma \in \{0, 8, 16\}$) to all outputs (Real, Step 2, and Step 50), progressively filtering out texture, as shown in Fig. 3(a). We then computed the inter-frame phase difference via 2D FFT to capture how the temporal phase structure evolves between consecutive frames. The alignment of these dynamics was quantified using the Pearson correlation between the real and generated phase difference maps. Even under strong blur ($\sigma = 16$), the Step 2 latent maintained a 3.6× higher correlation (0.358) compared to the Step 50 (0.100). This confirms that the phase consistency of Step 2 is a genuine structural property, rather than merely an artifact of blurriness.

Phase Sensitivity of Motion Dynamics. The above analysis establishes that phase degradation correlates with physical inconsistency. To confirm this relationship is causal, we isolate the effect of phase versus magnitude through controlled corruption experiments on real-world videos. Specifically, we decompose video frames via FFT and selectively inject 50% uniform noise into either the phase or magnitude spectrum while keeping the other intact. We then measure motion distortion using optical flow estimated by RAFT (Teed & Deng, 2020), quantified by End-Point Error (EPE), the average Euclidean distance (in pixels) between the original and corrupted flow vectors. As shown in Fig. 3(b), phase corruption induces severe motion distortion (EPE: 9.74, i.e., ∼10 pixel average displacement error), whereas equivalent magnitude corruption preserves motion accuracy (EPE: 1.14, ∼1 pixel error). This 8.5× disparity confirms that motion dynamics are intrinsically encoded in phase, not merely correlated but causally dependent. Additional causal studies on other metrics and details of the experimental setup are provided in the §Appendix B.2. Combined with our observation that denoising erodes phase by 18%, this explains 2-step videos retain better phase information, thereby preserving physical consistency.

4. PhaseLock

Building on our analysis that the few inference prior retains structural phase information, we propose **PhaseLock**, a training-free strategy that locks the phase information of the high-fidelity generation. While a direct substitution of the phase spectrum or a selective injection of low-frequency bands might seem intuitive, we avoid such explicit spectral manipulations as they often induce high-frequency artifacts and feature incoherence (See §Appendix D.3) Instead, PhaseLock introduces a spatial-domain proxy that locks phase dynamics by constraining the latent delta, a strategy we theoretically justify in Sec. 4.4. Our framework operates in two distinct stages: (1) **Motion Prior Extraction**, obtaining a structural guide from few inference steps (Sec. 4.2), and (2) **Latent Delta Guidance**, aligning the high-fidelity generation with the extracted motion priors (Sec. 4.3). A concise summary of the algorithm is available in the §Appendix Alg. 1.

4.1. Preliminaries

Let $\mathbf{x} \in \mathbb{R}^{F\times C\times H\times W}$ denote a video sequence with F frames. In Latent Diffusion Models (LDMs), a pre-trained VAE encoder $\mathcal{E}$ maps pixel-space frames to a latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{F\times C'\times H'\times W'}$. The diffusion process is defined by a forward chain that progressively adds Gaussian noise, and a reverse chain parameterized by a denoising model $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$, where t is the timestep and $\mathbf{c}$ denotes conditioning (*e.g.,* text and reference image).

Latent Delta Operator. We define the Latent Delta Operator $\mathcal{T}$ as the inter-frame difference in the latent space: $\mathcal{T}(\mathbf{z}) = \mathbf{z}_{2:F} - \mathbf{z}_{1:F-1}$, where $\mathbf{z}_{f:F}$ denotes the temporal slicing of the tensor from frame f to F. This operator captures local temporal dynamics while being invariant to

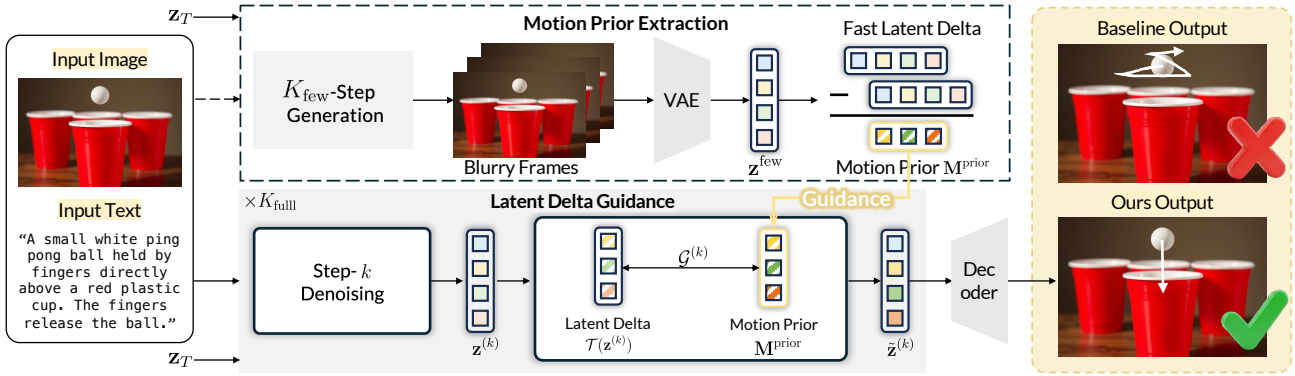


Figure 4. **The overall pipeline of PhaseLock.** Our method operates in two distinct stages. (1) Motion Prior Extraction: We derive frame-wise motion dynamics from a few inference trajectory. (2) Latent Delta Guidance: We transfer this motion prior into the standard denoising process. This training-free mechanism effectively enhance physical consistency while preserving high visual fidelity.

time-independent features (*e.g.,* static background).

4.2. Motion Prior Extraction

Guided by our analysis, we perform a few inference process to obtain the motion prior. Given a Gaussian noise initialization $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we generate a coarse latent sequence $\mathbf{z}^{\text{few}}$ using only K_{few} denoising steps (*e.g.,* $K_{\text{few}} = 2$): $\mathbf{z}^{\text{few}} = \mathcal{S}(\mathbf{z}_T, \mathbf{c}, K_{\text{few}}; \boldsymbol{\epsilon}_\theta)$, where $\mathbf{c}$ denotes the conditioning signals (*e.g.,* text prompts). Here, $\mathcal{S}$ represents the deterministic sampling process that iteratively maps the Gaussian noise to the data manifold using the pre-trained model $\boldsymbol{\epsilon}_\theta$. Note that we utilize the diffusion backbone $\boldsymbol{\epsilon}_\theta$ in a frozen state, requiring no additional fine-tuning.

Although latent magnitudes naturally fluctuate during denoising, our empirical analysis in Sec. 3 demonstrates that motion perception is robust to magnitude variations but highly sensitive to phase disruptions. This validates our use of the spatial difference as a phase-sensitive proxy to lock motion dynamics; we provide theoretical grounding in Sec. 4.4. We extract the motion template $\mathbf{M}^{\text{prior}}$ by applying the latent delta operator to the guide latent:

$$\mathbf{M}^{\text{prior}} = \mathcal{T}(\mathbf{z}^{\text{few}}) = \mathbf{z}^{\text{few}}_{2:F} - \mathbf{z}^{\text{few}}_{1:F-1}. \tag{1}$$

This tensor serves as the structural prior of physically plausible motion, which guides the full generation process to maintain dynamic consistency.

4.3. Latent Delta Guidance

In the second stage, we perform the standard high-fidelity generation with K_{full} steps (*e.g.,* $K_{\text{full}} = 50$). To ensure alignment with the motion prior, we re-initialize the noise $\mathbf{z}_T$ using the same seed s used in Stage 1. At each denoising step k, let $\mathbf{z}^{(k)}$ denote the current intermediate latent. We first compute the current motion dynamics $\mathbf{M}^{(k)} = \mathcal{T}(\mathbf{z}^{(k)})$. The guidance signal $\mathcal{G}^{(k)}$ is defined as the residual between the target motion prior and the current dynamics, $\mathcal{G}^{(k)} = \mathbf{M}^{\text{prior}} - \mathcal{T}(\mathbf{z}^{(k)})$. This signal captures the deviation of the current trajectory. We inject this guidance specifically into the subsequent frames to align their temporal evolution, while keeping the first frame (anchor) intact:

$$\mathbf{z}^{(k)}_{2:F} \leftarrow \mathbf{z}^{(k)}_{2:F} + \lambda(k) \cdot \mathcal{G}^{(k)}, \tag{2}$$

where $\lambda(k)$ is a time-dependent scalar controlling the guidance strength. Note that the first frame remains unmodified as it serves as the image condition anchor.

Adaptive Scheduling. Since the structure is determined in the early phase of diffusion, applying guidance in later steps may interfere with texture refinement. We explicitly decouple motion generation from detail refinement using a linear decay schedule. Let $k \in \{0, 1, \ldots, K_{\text{full}} - 1\}$ denote the elapsed denoising step (*i.e.,* $k = 0$ is the first step from pure noise). The guidance strength is active only during the interval $[k_{\text{start}}, k_{\text{end}})$:

$$\lambda(k) = \begin{cases} \lambda_0 \cdot \left(1 - \frac{k - k_{\text{start}}}{k_{\text{end}} - k_{\text{start}}}\right) & \text{if } k_{\text{start}} \le k < k_{\text{end}} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This schedule ensures strong adherence to the motion prior when the global layout is forming (early steps), and gradually relaxes constraints to allow for high-fidelity rendering.

4.4. Theoretical Justification

Building on our empirical findings, we provide theoretical insight into why constraining latent deltas effectively locks the phase evolution to the motion prior.

Latent Delta Encodes Phase Differences. Let $\mathcal{F}$ denote the Fourier transform. We analyze each frequency component separately; for a given frequency, let $A_f e^{j\phi_f}$ denote the corresponding Fourier coefficient of frame f, where A_f

Table 1. **Comprehensive evaluation on Physics-IQ.** We compare our training-free method against state-of-the-art video generation models. Our approach significantly improves the base models, surpassing much larger proprietary models.

Type	Model	Params	Score	Gain
Proprietary Models				
	Runway Gen-3 Alpha	-	22.8	-
	VideoPoet	-	20.3	-
	Lumiere	-	19.0	-
	Sora	-	10.0	-
Open-Source Models				
	MAGI-1	24B	30.2	-
	Stable Video Diffusion	3B	14.8	-
Ours (Plug-and-Play Integration)				
	CogVideoX (Base)	5B	30.8	-
	+ PhaseLock	5B	**36.0**	**(+5.2)**
	LTX-Video (Base)	2B	26.4	-
	+ PhaseLock	2B	32.0	**(+5.6)**
	WAN 2.1 (Base)	14B	20.9	-
	+ PhaseLock	14B	28.7	**(+7.8)**

Table 2. **Comparison of I2V models across PhyGenBench.** Images are generated using FLUX-schnell, unless marked with † (Gemini-3).

Model	Mech. (↑)	Optics (↑)	Therm. (↑)	Mat. (↑)	Avg. (↑)
CogVideoX	0.45	0.55	0.42	0.43	0.46
+ PhaseLock	0.51 (+13.3%)	0.78 (+41.8%)	0.47 (+11.9%)	0.49 (+14.0%)	0.57 (+23.9%)
CogVideoX†	0.53	0.78	0.45	0.48	0.51
+ PhaseLock	0.61(+15.1%)	0.80 (+2.6%)	0.49 (+8.9%)	0.52 (+8.3%)	0.61 (+19.6%)
WAN 2.1	0.43	0.55	0.38	0.30	0.42
+ PhaseLock	0.48 (+11.6%)	0.64 (+16.4%)	0.49 (+28.9%)	0.41 (+36.7%)	0.51 (+21.4%)
WAN 2.1†	0.45	0.60	0.41	0.32	0.46
+ PhaseLock	0.48 (+6.7%)	0.68 (+13.3%)	0.50 (+22.0%)	0.40 (+25.0%)	0.52 (+13.0%)

is the magnitude and ϕ_f is the phase. In natural video, consecutive frames typically share similar magnitude spectra, i.e., $A_f \approx A_{f-1} \triangleq A$. Under this assumption, the latent delta $\mathbf{\Delta} = \mathbf{z}_f - \mathbf{z}_{f-1}$ satisfies (full derivation in §Appendix C):

$$|\mathcal{F}(\mathbf{\Delta})| = 2A \left| \sin\left(\frac{\phi_f - \phi_{f-1}}{2}\right) \right| \approx A \cdot |\phi_f - \phi_{f-1}|, \quad (4)$$

where the approximation holds for small inter-frame phase shifts ($|\phi_f - \phi_{f-1}| \ll 1$), typical in smooth motion. This reveals that the latent delta magnitude is approximately proportional to the inter-frame phase difference, scaled by the shared magnitude A.

Connection to Motion Preservation. Since the latent delta encodes phase differences (Eq. 4), minimizing $\|\mathbf{M}^{\text{prior}} - \mathbf{M}^{(k)}\|$ directly constrains phase evolution—the component our causal ablation identifies as motion-critical (8.5× higher sensitivity than magnitude). Moreover, diffusion models are variance-preserving (Ho et al., 2022), ensuring magnitude variations do not dominate the loss even when $A_f \approx A_{f-1}$ is imperfect. This spatial-domain approach achieves phase alignment without explicit spectral manipulation that risks artifacts. Further detailed experiments and discussions, including ablation studies using spectral injection, are provided in §Appendix D.3.

5. Experiments

5.1. Setup

Baselines. To verify the universality and effectiveness of our proposed method, we conducted experiments on a diverse set of multiple video generation models, ranging from DiT-based architectures to recent open-source large-scale models. Specifically, we utilized CogVideoX-5B (Yang et al., 2024) as our primary baseline. Furthermore, we extended our evaluation to Wan (Wan et al., 2025), and LTX-Video (HaCohen et al., 2024) to demonstrate the robustness of our approach across different model architectures.

Implementation details. For the inference and evaluation pipeline, we utilized a single NVIDIA H100 (80GB) GPU. We adhered to the default hyperparameters provided by the official repositories of each backbone model unless otherwise specified. Specifically, CogVideoX-5B generates 49 frames at 8 fps with $K_{\text{full}} = 50$ steps; Wan produces 81 frames at 16 fps with $K_{\text{full}} = 50$ steps; LTX-Video outputs 121 frames at 30 fps with $K_{\text{full}} = 50$ steps. We applied our guidance mechanism without fine-tuning the pre-trained weights, ensuring a training-free adaptation. In our experiments, we set $\lambda_0 = 0.05$, $k_{\text{start}} = 0$, and $k_{\text{end}} = K_{\text{full}}/2$.

Evaluation. To comprehensively assess the performance of our method, we employed a multi-dimensional evaluation protocol covering both physical plausibility and general video quality. First, we measured physical accuracy using Physics-IQ (Motamed et al., 2025), which objectively calculates the kinematic deviation between generated trajectories and ground-truth simulations, and PhyGenBench (Meng et al., 2025), which evaluates perceptual physical plausibility through GPT-4o-based visual reasoning. Additionally, to ensure that the improvement in physical consistency does not compromise the overall visual fidelity, we reported VBench (Huang et al., 2024) scores.

5.2. Evaluation Results

Quantitative Evaluation: Physics-IQ Benchmark. To rigorously validate the physical consistency of our method, we conducted quantitative evaluations on the Physics-IQ benchmark (Motamed et al., 2025), which offers an objective measure of motion correctness beyond subjective visual metrics. We integrated PhaseLockinto diverse DiT architectures, including CogVideoX-5b, LTX-Video, and WAN, and observed a consistent performance boost, with an average

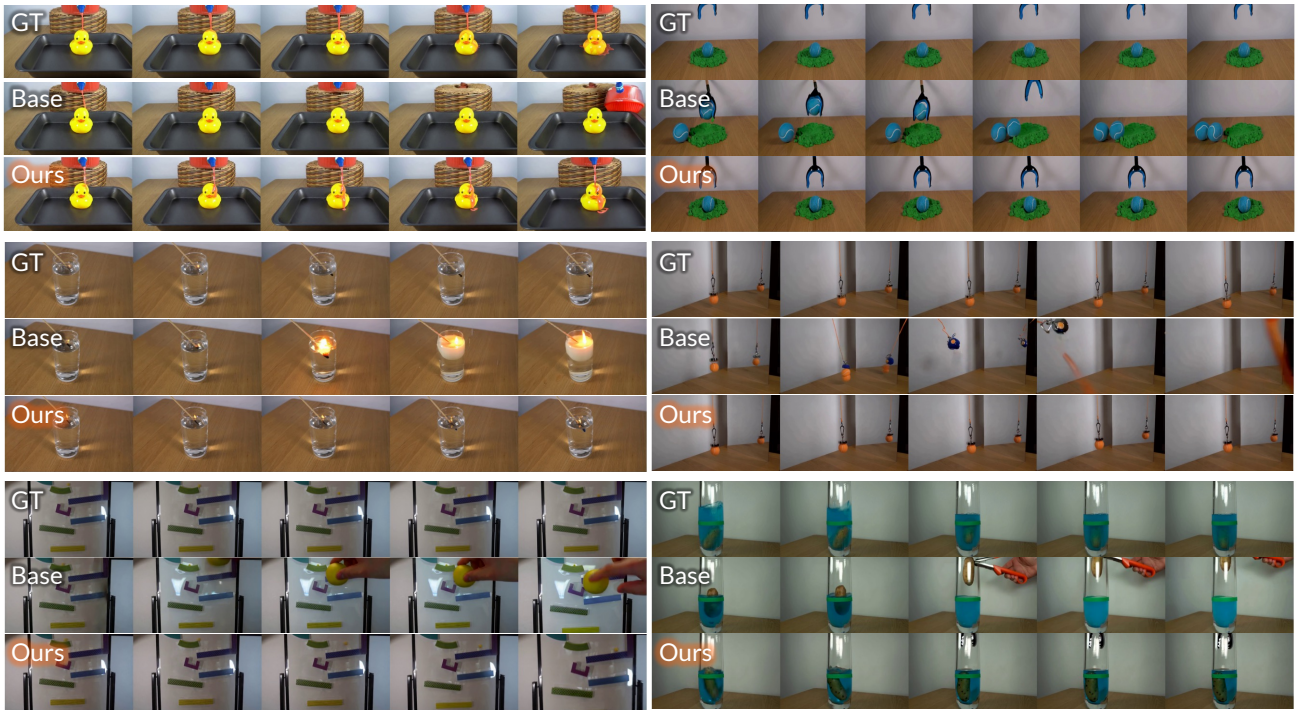


Figure 5. **Qualitative results on the Physics-IQ benchmark.** We compare the generated videos from the baseline ('Base') and our method ('Ours'). The results demonstrate that our method exhibits superior adherence to physical laws compared to the baseline, which often fails to maintain physical consistency.

Table 3. **Quantitative comparison using VBench.** Using VBench on the Physics-IQ and PhyGenBench benchmarks, we verified that our approach maintains visual fidelity, achieving scores that are either on par with or higher than the baselines.

Model	Subj. Cons.	Back. Cons.	Motion Smooth.	Temp. Flick.	Img. Qual.	Aesth. Qual.
CogVideoX-5B	**0.938**	0.940	**0.995**	0.994	0.664	0.467
+ PhaseLock	0.935	**0.955**	**0.995**	**0.995**	**0.680**	**0.489**
	(-0.3%)	(+1.5%)	(-0.0%)	(+0.1%)	(+2.3%)	(+4.7%)
WAN	**0.897**	0.911	0.987	0.983	0.643	**0.475**
+ PhaseLock	0.881	**0.921**	**0.995**	**0.993**	**0.655**	0.451
	(-1.8%)	(+1.0%)	(+0.7%)	(+0.9%)	(+1.8%)	(-5.2%)

increase of 6.2 % as shown in Table 1 (5.2% for CogVideoX, 5.6% for LTX-Video, 7.8% for WAN 2.1).

Quantitative Evaluation: PhyGenBench. Complementing our quantitative Physics-IQ, we employed PhyGenBench (Meng et al., 2025) to assess holistic physical plausibility via Large Vision Language Model (LVLM). To adapt this T2V benchmark for I2V, we utilized input images generated by Gemini-2.5-flash (Comanici et al., 2025) and FLUX-schnell (Esser et al., 2024) (details in §Appendix E). Despite the inherent subjectivity of LVLM evaluations, our method demonstrates superior physical robustness, outperforming baselines as shown in Table 2.

Table 4. **Human preference evaluation.** Pairwise comparisons against CogVideoX and WAN 2.1. We report both Win Rate (Win) and Accuracy (Acc) for each model. Ours consistently outperforms the baselines.

Criteria	Ours vs. CogVideoX				Ours vs. WAN 2.1			
	Ours		CogVideoX		Ours		WAN 2.1	
	Win	Acc	Win	Acc	Win	Acc	Win	Acc
Physics Plausibility	**72.9**	**58.0**	27.3	42.0	**81.9**	**67.1**	18.1	32.9
Visual Quality	**77.5**	**60.8**	22.5	39.2	**88.9**	**72.1**	11.1	27.9
Prompt Alignment	**62.6**	**54.1**	37.4	45.9	**78.4**	**63.1**	21.9	36.9

Quantitative Evaluation: VBench. We further verify that our physics improvements do not degrade general visual quality by evaluating generated Physics-IQ and PhyGenBench videos on VBench (Huang et al., 2024) across six visual quality dimensions. As shown in Table 3, our method preserves or improves most metrics across both CogVideoX-5B and WAN 2.1, with notable gains in Background Consistency (+1.5%, +1.0%) and Image Quality (+2.3%, +1.8%).

Quantitative Evaluation: Human Study. We also supplement our quantitative evaluations with a human study to verify the effectiveness of our approach, following the evaluation protocol of (Yuan et al., 2026). We conduct a human study on full videos of Physics-IQ benchmark with a side-by-side comparison interface where five annotators view

Figure 6. **Comparison of Efficiency and Performance.** Note that N denotes the number of generated samples. Our method achieves significant performance gains while maintaining low latency and memory usage comparable to the baseline. In contrast, achieving similar gains with other models requires more than $N\times$ the time and memory.

pairs of generated videos. For each video pair, annotators provide judgments across three criteria: Physics Plausibility (whether the motion follows realistic physical dynamics), Visual Quality (overall perceptual quality and clarity), and Motion Coherence (whether the object trajectory is smooth and consistent). Results are aggregated using win rates (excluding neutrals) and accuracy scores to account for ties $(\text{wins} + 0.5 \times \text{neutrals})/\text{total}$. Table 4 demonstrates that our method delivers significant improvement in all categories. This aligns with our quantitative findings that phase preservation primarily benefits physical motion dynamics rather than perceptual appearance. Further details in §Appendix E.

Qualitative Evaluation. Our method demonstrates the capability to synthesize dynamic elements as shown in Fig. 5. Compared to the baseline, our results are significantly more stable, effectively avoiding artifacts such as motion jitter and the random appearance of new objects. Specific examples highlight this physical consistency. In the bottom-right of Fig. 5, the liquid level naturally rises upon object entry, consistent with Archimedes' principle. In the bottom-left example, while the baseline fails to preserve small objects, causing the ball to vanish, our model maintains a smooth, continuous trajectory. See §Appendix F.2 for more results.

5.3. Efficiency & Trade-off Analysis

We analyze the computational overhead of PhaseLock in comparison to existing physics-alignment strategies. Fig. 6 presents a comparison with vLDM which aligns with our models. As illustrated in the graph, vLDM is reported to achieve the performance gains shown at $N = 16$. As shown in Fig. 6, our method incurs a minimal overhead of approximately 6% in inference time, primarily attributed to the truncated forward passes in extracting motion prior. The Latent Delta Injection involves lightweight latent tensor manipulations, which impose negligible latency compared to the heavy attention computations of the diffusion backbone. Notably, PhaseLock achieves a competitive Physics-IQ score of 36.0, closely matching the state-of-the-art reward-based method, WMReward (Yuan et al., 2026) (36.3). However, unlike WMReward, which relies on computationally intensive test-time optimization (gradient backpropagation), external Video-JEPA models, or expensive Best-of-N sampling ($\times N$ cost), our approach remains strictly training-free and gradient-free. This demonstrates that PhaseLock achieves performance comparable to state-of-the-art methods while maintaining superior computational efficiency.

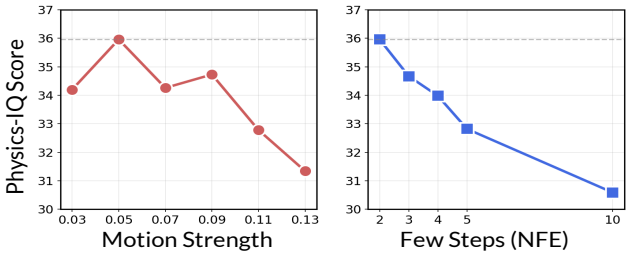


Figure 7. **Ablation studies on hyperparameters.** Impact of motion strength and number of few inference steps (NFE) on Physics-IQ scores. The performance peaks at strength 0.05 and highest score at few step NFE = 2.

5.4. Ablation Studies

To assess the sensitivity of our model to key hyperparameters, we conducted ablation studies on motion strength λ_0 and the number of few steps (NFE). First, regarding the motion strength, we varied the value from 0.03 to 0.13. As shown in Fig. 7, a strength of 0.05 yields the highest Physics-IQ score, whereas higher values lead to a decline in performance. Second, for the few inference steps, we evaluated NFE values ranging from 2 to 10. We observed that performance is maximized at NFE = 2, with a consistent decrease in scores as the number of steps increases. More ablation studies on scheduling, guidance formulation, and hyperparameters are provided in §Appendix F.1.

6. Conclusion

We reveal that video diffusion models establish physically consistent motion within just 2 steps, yet progressively overwrite this knowledge during visual refinement. PhaseLock locks these motion priors before they are lost, improving physical consistency by 6.2% on average at negligible cost ($1.06\times$ time) without external guidance or additional training [1]. Our findings suggest that physics-aware generation may require not more computation, but smarter preservation, opening directions toward phase, preserving samplers, physics-aware training objectives, and extensions to audio and 3D domains.

[1] https://phaselock-physical-video.github.io/

Impact Statement

Our research investigates the internal mechanisms of diffusion models to mitigate physical hallucinations in video generation. By ensuring generated content adheres to physical laws, this work contributes to applications requiring high reliability, such as robotics simulation and scientific visualization. We acknowledge that advancing generative capabilities implies potential societal risks, including the creation of misleading media. However, we believe that understanding the limitations and mechanisms of these models is essential for building more transparent, controllable, and physically grounded AI systems.

References

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., and Goldstein, T. Cold diffusion: Inverting arbitrary image transforms without noise. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.-W., and Grover, A. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al. Lumiere: A space-time diffusion model for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorber, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. In *arXiv preprint arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, 2023b.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al. Video generation models as world simulators. *OpenAI Technical Report*, 2024.

Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481, 2022.

Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., and Wang, Y. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dieleman, S. Diffusion is spectral autoregression. *Blog post*, 2024. URL https://sander.ai/2024/09/02/spectral-autoregression.html.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Gao, Q., Pi, X., Liu, K., et al. Do vision-language models have internal world models? towards an atomic evaluation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26170–26195, 2025.

Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S. S., Shah, A., Yin, X., Parikh, D., and Misra, I. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., and Feng, J. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.-C., et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Langley, P. Crafting papers on machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 1207–1216, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Li, L., Xu, J., Dong, Q., Zheng, C., Sun, X., Kong, L., and Liu, Q. Can language models understand physical concepts? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11843–11861, 2023.

Li, Z., Wu, X., Shi, G., Qin, Y., Du, H., Liu, F., Zhou, T., Manocha, D., and Boyd-Graber, J. L. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Liu, S., Wang, Z., Pirk, S., Savarese, S., Wu, J., et al. Physgen: Rigid-body physics-grounded image-to-video generation. *European Conference on Computer Vision*, 2024.

Meng, F., Liao, J., Tan, X., Lu, Q., Shao, W., Zhang, K., Cheng, Y., Li, D., and Luo, P. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 43781–43806, 2025.

Motamed, S., Culp, L., Swersky, K., Jaini, P., and Geirhos, R. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. *Discrete-Time Signal Processing*. Prentice Hall, 2nd edition, 1999.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Piloto, L. S., Weinstein, A., Battaglia, P., et al. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6 (9):1257–1267, 2022.

Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

Si, C., Huang, Z., Jiang, Y., and Liu, Z. Freeu: Free lunch in diffusion u-net. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4733–4743, 2024.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Teed, Z. and Deng, J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision (ECCV)*, pp. 402–419, 2020.

Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025.

Wu, T., Si, C., Jiang, Y., Huang, Z., and Liu, Z. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, 2024.

Xue, Q., Yin, X., Yang, B., and Gao, W. PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18826–18836, June 2025.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Yuan, J., Zhang, X., Friedrich, F., Beltran-Velez, N., Hall, M., Askari-Hemmat, R., Han, X., Ballas, N., Drozdzal, M., and Romero-Soriano, A. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026.

Yuan, Y., Song, J., Iqbal, U., Vahdat, A., and Kautz, J. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16010–16021, 2023.

Yuan, Y., Guo, Y., Wang, C., Zhang, W., Xu, H., and Zhang, L. Freqprior: Improving video diffusion models with frequency filtering gaussian noise. *arXiv preprint arXiv:2502.03496*, 2025.

Zhang, K., Xiao, C., Xu, J., Mei, Y., and Patel, V. M. Think before you diffuse: Infusing physical rules into video diffusion. *arXiv preprint arXiv:2505.21653*, 2025a.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 586–595, 2018.

Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., and Cheng, Y. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025b.

Technical Appendices

A. Additional Material: Media Gallery

As displaying video content frame by frame within the paper offers only limited insight into temporal coherence and visual quality, we provide a Media Gallery Page featuring the full video outputs from both the main and additional experiments. This page allows for a more faithful assessment of motion consistency and prompt alignment. The results can be viewed at: https://phaselock-physical-video.github.io/

B. Analyses Details

B.1. Setup Details

To conduct the analysis for our main paper, we utilized the Physics-IQ benchmark (Motamed et al., 2025), which provides Ground Truth (GT) videos. We employed Diffusion Transformer models, specifically CogVideoX (Yang et al., 2024) and WAN 2.1 (Wan et al., 2025). The analysis was performed using 30 to 50 randomly selected paired samples. Our analyze and method relies on the denoising process; therefore, Autoregressive Transformer-based models (Yuan et al., 2026) are not applicable to this study. All experiments were conducted on NVIDIA H100 (80GB) GPUs. Unless explicitly stated otherwise, we adopted the default hyperparameters provided by the official repositories. We applied the benchmark settings, using identical input modalities, including text prompts and reference images.

Divergence of Visual and Physical Fidelity. In Fig 1 at main paper, we present a quantitative comparison between the baseline (at denoising steps $t = \{2, 50\}$) and our proposed method. We utilized the comprehensive Physics-IQ benchmark to objectively measure physical consistency, specifically quantifying frame-by-frame motion dynamics (Physics Score). For visual fidelity, we employed the LPIPS metric (Zhang et al., 2018). To facilitate an intuitive visualization where higher values indicate better performance for both axes, we report visual quality as '$1 - \text{LPIPS}$'. Complementing the figure, Table 5 provides the detailed numerical results, including additional evaluations for the baseline at intermediate steps ($t = 10, 30$).

Table 5. **Quantitative comparison on the Physics-IQ benchmark.** We evaluate the physical consistency and visual quality across different denoising steps of the baseline and our method PHASELOCK. Note that Visual Quality is reported as $(1 - \text{LPIPS})$ so that higher values indicate better quality for both metrics.

Method	**Physics-IQ** (↑)	**Visual Quality** ($1 - \text{LPIPS}$, ↑)
CogVideoX (Step 2)	34.02	0.7721
CogVideoX (Step 10)	32.84	0.8183
CogVideoX (Step 30)	31.76	0.8022
CogVideoX (Step 50)	30.32	0.8073
Ours	**36.00**	0.8020

Spatio-Temporal Slicing. To explicitly analyze the motion trajectories, we employed spatio-temporal slicing (often referred to as x-t slices). Specifically, we defined a cross-section indicated by a yellow reference line on the video frames and extracted the temporal evolution of pixels along this axis. We visualized the trajectories at Step 2 and Step 50 using samples from the Physics-IQ benchmark in Fig. 16, Fig. 17, and Fig. 18. This visualization allows us to intuitively assess the continuity and physical plausibility of the generated motion across different inference steps.

B.2. Causal Studies

To causally verify the hypothesis that phase information is the primary carrier of physical dynamics, we conducted a controlled causal study utilizing CogVideoX as a representative model. We systematically injected synthetic noise into the phase and magnitude spectra, varying the corruption level from $\alpha = 0$ (original) to $\alpha = 1$ (fully corrupted) for each component separately and quantified the resulting physical degradation across three hierarchical levels of motion derivatives: (1) object trajectory (position), (2) velocity profile (speed), and (3) motion smoothness (jerk). All quantitative evaluations were performed by comparing the generated outputs against the GT videos.

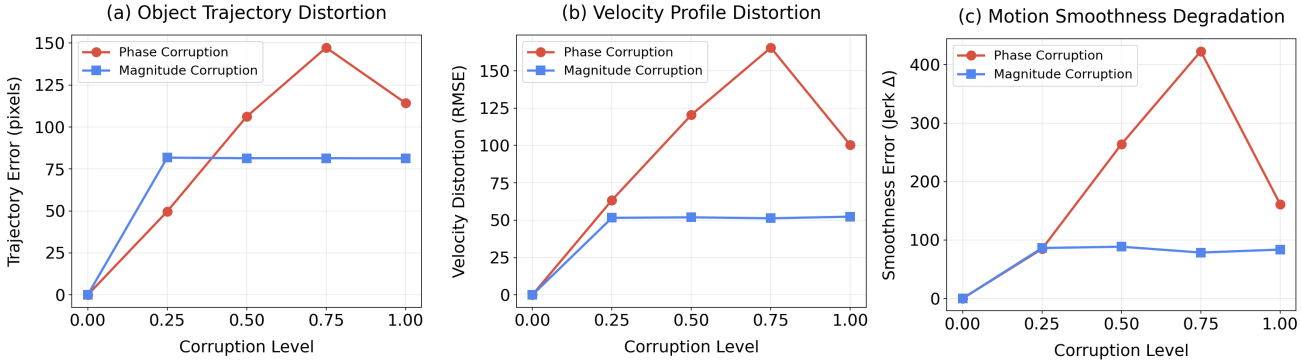


Figure 8. **Causal analysis of phase and magnitude corruption on physical dynamics.** We compare the impact of spectral corruption (α) across three hierarchical metrics. *(a) Trajectory Error (Position):* Phase corruption causes continuous drift in object location, while magnitude error saturates at $\alpha = 0.25$. *(b) Velocity Distortion (Speed):* Phase corruption severely impacts velocity consistency compared to the invariant response of magnitude. *(c) Motion Smoothness (Jerk):* Phase noise induces physically impossible, jerky motion, showing explosive error growth in higher-order derivatives.

(1) Object Trajectory Error. This metric measures the positional deviation of the object, answering "Is the object where it is supposed to be?" We extract the object centroid $\mathbf{c} = (c_x, c_y)$ using image moments M_{pq} derived from Canny edge maps:

$$c_x = \frac{M_{10}}{M_{00}}, \quad c_y = \frac{M_{01}}{M_{00}} \tag{5}$$

The trajectory error is defined as the Euclidean distance between the original and corrupted centroids over time T:

$$E_{\text{traj}} = \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{p}_t^{\text{orig}} - \mathbf{p}_t^{\text{corr}}\|_2 \tag{6}$$

As shown in Fig. 8(a), magnitude corruption leads to an immediate saturation in error (≈ 81px at $\alpha = 0.25$) but does not degrade further, as the structural edges remain intact despite visual artifacts (*e.g.,* blurring or contrast shifts). In contrast, phase corruption causes a progressive increase in error, eventually surpassing magnitude error at $\alpha \geq 0.5$. This confirms that phase encodes the structural location of the object.

(2) Velocity Profile Distortion. This metric evaluates whether the object follows physical laws of motion, such as the linear velocity increase in free-fall ($v \propto t$). We compute the velocity sequence $v_t = \|\mathbf{p}_{t+1} - \mathbf{p}_t\|_2$ and measure the Root Mean Square Error (RMSE):

$$E_{\text{vel}} = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T-1} (v_t^{\text{orig}} - v_t^{\text{corr}})^2} \tag{7}$$

As shown in Fig. 8 (b), magnitude corruption shows negligible impact on velocity profiles, with error remaining constant regardless of noise intensity. However, phase corruption induces severe distortion, resulting in $2.32\times$ higher error than magnitude at $\alpha = 0.5$. Since velocity is the first derivative of position ($v = d\mathbf{p}/dt$), small positional jitters caused by phase noise are amplified, destroying the linear consistency required for physical realism (*e.g.,* constant acceleration).

(3) Motion Smoothness Error. To assess the naturalness of the motion, we analyze the Jerk (the rate of change of acceleration). In natural physics (*e.g.,* gravity), acceleration is constant ($\mathbf{a} \approx g$), implying that jerk should be minimal ($\mathbf{j} \approx 0$). We define the jerk vector $\mathbf{j}_t$ and the smoothness error E_{smooth} as:

$$\mathbf{j}_t = \mathbf{p}_{t+3} - 3\mathbf{p}_{t+2} + 3\mathbf{p}_{t+1} - \mathbf{p}_t, \quad E_{\text{smooth}} = |\bar{J}^{\text{corr}} - \bar{J}^{\text{orig}}| \tag{8}$$

The impact of phase corruption becomes most pronounced in high-order derivatives. While magnitude corruption has almost no effect on smoothness (ratio ≈ 1.0), phase corruption causes an explosive increase in jerk error, reaching $5.39\times$ the error of magnitude at $\alpha = 0.75$ as shown in Fig. 8(c). This demonstrates that phase noise manifests as physically impossible, jerky motion.

C. Proof of Latent Delta Encoding Phase Differences

We provide the full mathematical derivation of Eq. 4 in the main paper, which establishes that the latent delta encodes inter-frame phase differences.

C.1. Setup and Notation

Let $\mathcal{F}$ denote the Fourier transform. For a given frequency component, let the Fourier coefficients of consecutive frames $f-1$ and f be:

$$\mathcal{F}(z_{f-1}) = A_{f-1} \cdot e^{j\phi_{f-1}}, \tag{9}$$

$$\mathcal{F}(z_f) = A_f \cdot e^{j\phi_f}, \tag{10}$$

where $A_f, A_{f-1} \in \mathbb{R}^+$ are magnitudes and $\phi_f, \phi_{f-1} \in [-\pi, \pi)$ are phases.

C.2. Derivation

Proof. The latent delta is defined as $\mathbf{\Delta} = z_f - z_{f-1}$. By linearity of the Fourier transform:

$$\mathcal{F}(\mathbf{\Delta}) = \mathcal{F}(z_f) - \mathcal{F}(z_{f-1}) = A_f e^{j\phi_f} - A_{f-1} e^{j\phi_{f-1}}. \tag{11}$$

Step 1: Magnitude Stability Assumption. In natural video, consecutive frames exhibit similar magnitude spectra due to temporal continuity:

$$A_f \approx A_{f-1} \triangleq A. \tag{12}$$

Under this assumption:

$$\mathcal{F}(\mathbf{\Delta}) = A\left(e^{j\phi_f} - e^{j\phi_{f-1}}\right). \tag{13}$$

Step 2: Complex Exponential Difference. We simplify $e^{j\phi_f} - e^{j\phi_{f-1}}$ using the identity for difference of complex exponentials. Define:

$$\bar{\phi} = \frac{\phi_f + \phi_{f-1}}{2}, \quad \Delta\phi = \phi_f - \phi_{f-1}. \tag{14}$$

Then $\phi_f = \bar{\phi} + \frac{\Delta\phi}{2}$ and $\phi_{f-1} = \bar{\phi} - \frac{\Delta\phi}{2}$. Substituting:

$$e^{j\phi_f} - e^{j\phi_{f-1}} = e^{j(\bar{\phi}+\Delta\phi/2)} - e^{j(\bar{\phi}-\Delta\phi/2)} \tag{15}$$

$$= e^{j\bar{\phi}}\left(e^{j\Delta\phi/2} - e^{-j\Delta\phi/2}\right). \tag{16}$$

Step 3: Euler's Formula. Using Euler's formula, $e^{jx} - e^{-jx} = 2j\sin(x)$:

$$e^{j\phi_f} - e^{j\phi_{f-1}} = e^{j\bar{\phi}} \cdot 2j \sin\left(\frac{\Delta\phi}{2}\right). \tag{17}$$

Step 4: Taking the Magnitude. Since $|e^{j\bar{\phi}}| = 1$ and $|j| = 1$:

$$\left|e^{j\phi_f} - e^{j\phi_{f-1}}\right| = 2\left|\sin\left(\frac{\Delta\phi}{2}\right)\right| = 2\left|\sin\left(\frac{\phi_f - \phi_{f-1}}{2}\right)\right|. \tag{18}$$

Therefore:

$$\boxed{|\mathcal{F}(\mathbf{\Delta})| = 2A\left|\sin\left(\frac{\phi_f - \phi_{f-1}}{2}\right)\right|.} \tag{19}$$

Step 5: Small Angle Approximation. For smooth motion, inter-frame phase shifts are small: $|\phi_f - \phi_{f-1}| \ll 1$. Using Taylor expansion $\sin(x) \approx x$ for $|x| \ll 1$:

$$|\mathcal{F}(\mathbf{\Delta})| \approx 2A \cdot \frac{|\phi_f - \phi_{f-1}|}{2} = A \cdot |\phi_f - \phi_{f-1}|. \tag{20}$$

□

C.3. Interpretation

This result establishes that:

$$|\mathcal{F}(\mathbf{\Delta})| \propto |\phi_f - \phi_{f-1}|, \tag{21}$$

i.e., the latent delta magnitude is directly proportional to the inter-frame phase difference. Consequently, minimizing the latent delta error $\|\mathbf{M}^{\text{prior}} - \mathbf{M}^{(k)}\|$ in the spatial domain effectively constrains the phase evolution in the frequency domain, locking it to the motion prior.

C.4. General Case: Unequal Magnitudes

When the magnitude stability assumption does not hold exactly ($A_f \neq A_{f-1}$), we derive the general expression.

Proof. Starting from:

$$\mathcal{F}(\mathbf{\Delta}) = A_f e^{j\phi_f} - A_{f-1} e^{j\phi_{f-1}}. \tag{22}$$

The squared magnitude is:

$$\begin{aligned}
|\mathcal{F}(\mathbf{\Delta})|^2 &= \mathcal{F}(\mathbf{\Delta}) \cdot \overline{\mathcal{F}(\mathbf{\Delta})} && (23)\\
&= \left(A_f e^{j\phi_f} - A_{f-1} e^{j\phi_{f-1}}\right)\left(A_f e^{-j\phi_f} - A_{f-1} e^{-j\phi_{f-1}}\right) && (24)\\
&= A_f^2 + A_{f-1}^2 - A_f A_{f-1} e^{j(\phi_f - \phi_{f-1})} - A_f A_{f-1} e^{-j(\phi_f - \phi_{f-1})} && (25)\\
&= A_f^2 + A_{f-1}^2 - 2A_f A_{f-1} \cos(\phi_f - \phi_{f-1}). && (26)
\end{aligned}$$

Therefore:

$$\boxed{|\mathcal{F}(\mathbf{\Delta})|^2 = A_f^2 + A_{f-1}^2 - 2A_f A_{f-1} \cos(\phi_f - \phi_{f-1}).} \tag{27}$$

□

This is the law of cosines for complex vectors. The phase difference $(\phi_f - \phi_{f-1})$ controls the cosine term, which provides the primary modulation. Even with magnitude variations, the phase difference remains the dominant factor determining $|\mathcal{F}(\mathbf{\Delta})|$.

When $A_f = A_{f-1} = A$:

$$|\mathcal{F}(\mathbf{\Delta})|^2 = 2A^2(1 - \cos(\phi_f - \phi_{f-1})) = 4A^2 \sin^2\left(\frac{\phi_f - \phi_{f-1}}{2}\right), \tag{28}$$

which recovers our earlier result.

D. Methods Details

D.1. Algorithm of PhaseLock

We present the detailed inference pseudocode for PhaseLockin Algorithm 1. The process consists of two stages: (1) extracting a coarse motion prior ($\mathbf{M}^{\text{prior}}$) using a few-step sampling strategy, and (2) guiding the full generation process to align with this prior via a linearly decaying guidance strength λ.

Algorithm 1 PHASELOCK

Require: Reference Image $\mathbf{x}_{\text{ref}}$, Text Prompt p, Pre-trained Diffusion Model $\boldsymbol{\epsilon}_\theta$
Require: few Steps $K_{\text{few}} = 2$, Full Steps $K_{\text{full}} = 50$
Require: Guidance Interval $[k_{\text{start}}, k_{\text{end}})$, Strength λ_0
// (1) Motion Prior Extraction
$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{z}^{\text{few}} \leftarrow \mathcal{S}(\mathbf{z}_T, \mathbf{c}, K_{\text{few}}; \boldsymbol{\epsilon}_\theta)$,
$\mathbf{M}^{\text{prior}} \leftarrow \mathbf{z}^{\text{few}}_{2:F} - \mathbf{z}^{\text{few}}_{1:F-1}$ {Extract temporal deltas}
// (2) Guided Generation
$\mathbf{z} \leftarrow \mathbf{z}_T$ {Re-use initial noise}
for $k = 0$ to $K_{\text{full}} - 1$ **do**
  $\mathbf{z} \leftarrow \text{Step}(\mathbf{z}, \boldsymbol{\epsilon}_\theta, \mathbf{x}_{ref}, p)$ {One denoising step}
  // Apply Guidance
  if $k_{\text{start}} \leq k < k_{\text{end}}$ **then**
    $\mathbf{M} \leftarrow \mathbf{z}_{2:F} - \mathbf{z}_{1:F-1}$ {Current motion}
    $\mathcal{G} \leftarrow \mathbf{M}^{\text{prior}} - \mathbf{M}$
    $\lambda \leftarrow \lambda_0 \cdot \left(1 - \frac{k - k_{\text{start}}}{k_{\text{end}} - k_{\text{start}}}\right)$ {Linear decay}
    $\mathbf{z}_{2:F} \leftarrow \mathbf{z}_{2:F} + \lambda \cdot \mathcal{G}$
  end if
end for
return Decode$(\mathbf{z})$

D.2. Impact of Inference Steps on Physical Consistency

We conducted a step-wise analysis using the CogVideoX-5b (I2V) model to address concerns that our results might be attributed to additional processing steps. Although higher NFEs (Number of Function Evaluations) are often associated with better quality, our experiments show that this does not guarantee improved physical accuracy and comes with high computational overhead. Table 6 illustrates that even when the number of timesteps is doubled, the model shows only marginal gain of about 1%, remaining significantly lower than our method's performance of 36.0%.

Table 6. **Quantitative comparison of Physics-IQ scores across varying inference steps (NFE)**

Method	NFE (timesteps)	Physics-IQ Score
Baseline (CogVideoX)	50	30.82
	51	31.44
	52	31.02
	53	31.54
	54	31.14
	55	30.74
	60	31.51
	80	31.94
	100	31.96
+ PHASELOCK	50 (+2)	36.0

D.3. Why Direct Frequency Manipulation Fails

A natural question arises: if phase information encodes motion dynamics, why not directly inject the low-frequency phase from the 2-step prior into the 50-step generation? In this section, we investigate several direct frequency manipulation baselines and explain why they fail to preserve physical consistency, motivating our spatial-domain Latent Delta Guidance approach.

D.3.1. BASELINE METHODS

We evaluate four direct frequency manipulation strategies:

(1) Low-Frequency Phase Injection. Extract the low-frequency phase from the 2-step prior and inject it into the 50-step latent at each denoising step:

$$\mathcal{F}(\mathbf{z}^{(k)}_{\text{guided}}) = |\mathcal{F}(\mathbf{z}^{(k)})| \cdot \exp\left(j \cdot \left[\mathbf{M}_{\text{LP}} \odot \phi^{\text{prior}} + (1 - \mathbf{M}_{\text{LP}}) \odot \phi^{(k)}\right]\right), \tag{29}$$

where $\mathbf{M}_{\text{LP}}$ is a Gaussian low-pass filter with cutoff frequency d_s (spatial) and d_t (temporal), and ϕ^{prior}, $\phi^{(k)}$ denote the phase spectra of the prior and current latent, respectively.

(2) Full Phase Substitution. Replace the entire phase spectrum with that of the 2-step prior while retaining the magnitude from the 50-step generation:

$$\mathbf{z}^{(k)}_{\text{guided}} = \mathcal{F}^{-1}\left(|\mathcal{F}(\mathbf{z}^{(k)})| \cdot \exp(j \cdot \phi^{\text{prior}})\right). \tag{30}$$

(3) Iterative Refinement. Following FreeInit (Wu et al., 2024), we iteratively refine the initial noise by preserving low-frequency components from the denoised latent and resampling high-frequency components:

$$\mathbf{z}_T^{(i+1)} = \mathcal{F}^{-1}\left(\mathbf{M}_{\text{LP}} \odot \mathcal{F}(\tilde{\mathbf{z}}_T^{(i)}) + (1 - \mathbf{M}_{\text{LP}}) \odot \mathcal{F}(\boldsymbol{\eta})\right), \tag{31}$$

where $\tilde{\mathbf{z}}_T^{(i)}$ is the re-noised latent from iteration i and $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I})$ is fresh Gaussian noise.

(4) Magnitude-Preserving Phase Blend. Linearly blend the phase spectra while preserving the original magnitude:

$$\phi_{\text{blend}} = \alpha \cdot \phi^{\text{prior}} + (1 - \alpha) \cdot \phi^{(k)}, \quad \mathbf{z}^{(k)}_{\text{guided}} = \mathcal{F}^{-1}\left(|\mathcal{F}(\mathbf{z}^{(k)})| \cdot \exp(j \cdot \phi_{\text{blend}})\right). \tag{32}$$

D.3.2. EXPERIMENTAL SETUP

We conduct experiments on CogVideoX-5B using the Physics-IQ benchmark. For all frequency manipulation methods, we apply guidance at each denoising step $k \in [0, K/2)$ to match the schedule of our proposed method.

Table 7 presents the experimental results comparing direct frequency manipulation methods against our approach.

Table 7. **Comparison of frequency manipulation baselines vs. our method on CogVideoX-5B.** All frequency methods perform catastrophically worse than baseline. They don't just fail to improve motion, they destroy it.

Method	Physics-IQ (↑)	Δ from Ours
CogVideoX Baseline (50-step)	30.90	-5.06
Magnitude-Preserving Phase Blend (α=0.5)	14.45	-21.51
Full Phase Substitution	14.21	-21.75
Low-Freq Phase Injection	13.69	-22.27
Iterative Refinement (n=2)	1.42	-34.54
Ours (Latent Delta)	**35.96**	—

All frequency manipulation methods perform catastrophically worse than even the unmodified baseline. These methods don't merely fail to improve physical consistency, they actively destroy the motion dynamics that the baseline model already captures.

D.3.3. ANALYSIS: WHY DIRECT FREQUENCY MANIPULATION FAILS

We identify three fundamental reasons why direct frequency manipulation fails to preserve physical consistency.

(1) Spectral Artifacts from FFT Operations. FFT-based manipulation introduces multiple sources of artifacts. First, the FFT assumes periodic boundary conditions, but video latents have no such periodicity; this mismatch causes *spectral leakage* at spatial and temporal boundaries. Second, frequency-domain filtering produces *ringing artifacts* due to the filter's impulse response, particularly at object boundaries and motion discontinuities where sharp transitions exist. These artifacts propagate through subsequent denoising steps and compound into visible distortions.

(2) Latent Space Structure Mismatch. Direct phase manipulation assumes that phase and magnitude form an appropriate decomposition for the learned latent space. However, VAE encoders learn representations optimized for reconstruction fidelity, not spectral separability. We hypothesize that the encoder learns a joint representation where phase and magnitude together encode semantic content in a coupled manner. Substituting phase while preserving magnitude creates hybrid representations that likely fall outside the learned latent manifold, producing outputs the decoder cannot properly reconstruct. This hypothesis is supported by our causal ablation (Sec. 3), which shows that motion is $8.5\times$ more sensitive to phase corruption than magnitude corruption. This extreme sensitivity suggests that phase carries critical structural information that cannot be surgically replaced without disrupting the latent's coherence.

(3) Frequency Domain $\neq$ Physical Property Preservation. Low-frequency components capture global appearance and coarse motion trajectories, but physical consistency involves constraints that span all frequencies:

- **Conservation laws**: Momentum and energy conservation operate across all spatial frequencies, not just low-frequency bands.
- **Collision dynamics**: Object collisions and contact events involve sharp discontinuities that manifest as high-frequency components.
- **Causal relationships**: Physical causality (cause precedes effect) is a temporal ordering constraint that is not localized to any frequency band.

FreeInit (Wu et al., 2024) improves *temporal smoothness* by preserving low-frequency temporal correlations, but temporal smoothness is neither necessary nor sufficient for physical correctness. A ball smoothly floating upward violates physics despite being temporally consistent, while a ball bouncing with sharp velocity reversals obeys physics despite temporal discontinuities.

D.3.4. Why Latent Delta Guidance Succeeds

Our Latent Delta Guidance avoids these failure modes through a fundamentally different approach: rather than surgically manipulating spectral components, we constrain the *distribution of inter-frame changes* in the spatial domain.

(1) Artifact-Free Operation. By operating entirely in the spatial domain, we never invoke FFT/IFFT operations on the latent representation. This eliminates spectral leakage from boundary discontinuities and ringing artifacts from frequency filtering.

(2) Aggregate Spectral Constraint via Parseval's Theorem. While we do not manipulate individual frequency components, Parseval's theorem provides an elegant connection between spatial and frequency domains:

$$\|\mathbf{M}^{\text{prior}} - \mathbf{M}^{(k)}\|^2 = \sum_{\omega} |\mathcal{F}(\mathbf{M}^{\text{prior}} - \mathbf{M}^{(k)})[\omega]|^2. \tag{33}$$

Minimizing the left-hand side in the spatial domain equivalently minimizes the sum of squared spectral differences. Combined with our analysis in Appendix C, which shows that latent delta magnitude is dominated by inter-frame phase differences (under typical video statistics), this provides an *aggregate constraint* on phase evolution without requiring per-frequency intervention.

(3) Magnitude-Weighted Prioritization. The aggregate constraint naturally weights frequency components by their squared magnitude $A[\omega]^2$. This is a desirable property: high-magnitude components carry the most perceptually significant motion information, while low-magnitude components (often corresponding to noise or fine texture) contribute minimally to the loss. The guidance thus focuses on dominant motion patterns while allowing flexibility in perceptually less important details.

(4) Respecting the Denoising Trajectory. Unlike spectral surgery that forces specific frequency values, our guidance signal $\mathcal{G}^{(k)} = \mathbf{M}^{\text{prior}} - \mathbf{M}^{(k)}$ operates as a soft constraint that nudges the denoising trajectory toward the motion prior. The diffusion model retains the ability to find a coherent solution within its learned manifold, rather than being forced into potentially inconsistent states.

Key Insight. The critical difference from direct phase manipulation is that we constrain *how latent representations change between frames*, not their absolute spectral values. This respects the diffusion model's learned dynamics while providing sufficient aggregate guidance to transfer motion priors. The empirical results (Table 7) validate that such aggregate constraints are not only sufficient but substantially more effective than per-frequency surgical intervention—precisely because they avoid disrupting the latent space structure that the model relies upon.

D.4. Spectral Analysis of PhaseLock

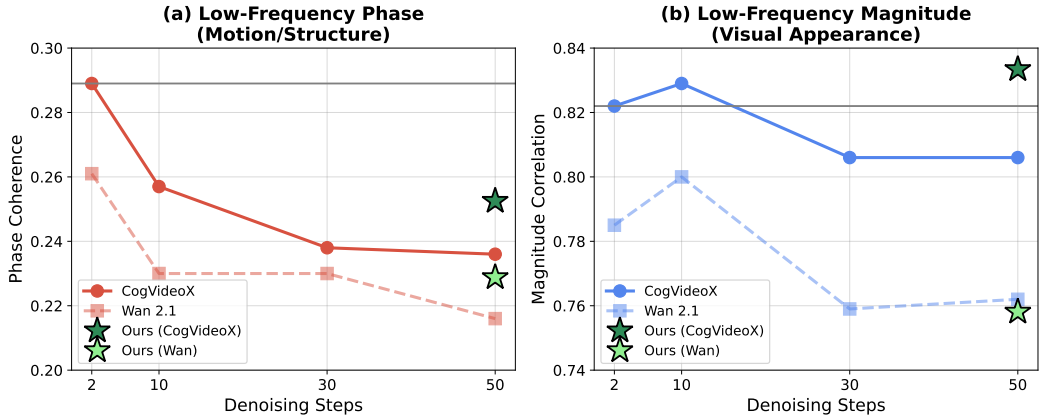


Figure 9. **Additional Analysis of PhaseLock.** Our Model, PhaseLock, mitigates phase erosion during denoising process without explicit FFT operations.

To quantitatively verify that our method preserves phase information as analyzed in Sec. 3, we additionally measured the low-frequency phase coherence and magnitude correlation between the videos generated by our method and the ground-truth real videos. We extract low-frequency components (within 40% of the Nyquist frequency) from the 3D FFT of 16-frame sequences, then compute phase coherence as $\mathbb{E}[\cos(\phi_{\text{real}} - \phi_{\text{gen}})]$ and magnitude correlation via Pearson's r on log-magnitude spectra. For baseline models, phase coherence degrades substantially as denoising progresses: CogVideoX drops from 0.289 (2-step) to 0.236 (50-step), an 18.3% relative decrease, while Wan exhibits a similar 17.2% drop (0.261 to 0.216). In contrast, magnitude correlation remains stable across all steps (<3% variation), confirming that phase, encoding object positions and motion direction, is the vulnerable component during denoising. Notably, our method at 50 steps achieves phase coherence of 0.253 (CogVideoX) and 0.229 (Wan), closely matching the *10-step* baseline levels (0.257 and 0.230) rather than the degraded 50-step values. This demonstrates that PHASELOCK successfully mitigates phase erosion without explicit FFT operations. By constraining frame deltas in the spatial domain, we indirectly preserve the inter-frame phase differences that encode motion dynamics (Eq. 4), achieving phase preservation while avoiding artifacts from direct spectral manipulation.

E. Evaluation Details

To objectively assess physical motion dynamics, we employed the Physics-IQ benchmark (Motamed et al., 2025). Additionally, we utilized PhyGenBench (Meng et al., 2025), which leverages Vision-Language Models (VLMs) for evaluation, as well as VBench (Huang et al., 2024). To mitigate the potential limitations of automated metrics and ensure a robust evaluation, we complemented these methods with a Human Study, strictly adhering to the protocols established in WMReward (Yuan et al., 2026). A detailed description of each evaluation methodology is provided below.

Physics-IQ Benchmark. To evaluate the physical understanding of video generation models, we utilized the Physics-IQ benchmark (Motamed et al., 2025). This benchmark comprises a diverse collection of 396 high-quality, real-world videos covering 66 distinct physical scenarios, ranging from fundamental principles such as solid mechanics and fluid dynamics

to complex phenomena like optics and magnetism. The evaluation protocol requires models to predict a 5 second video continuation based on initial conditioning frames. Performance is quantitatively assessed by comparing the generated motion against Ground Truth (GT) videos using a set of physics-aware metrics (including Spatial IoU, Spatiotemporal IoU, and Mean Squared Error (MSE)), which are aggregated into a unified Physics-IQ score to measure the model's adherence to real-world physical laws.

PhyGenBench Evaluation. We also incorporated PhyGenBench (Meng et al., 2025) to assess the model's grasp of physical commonsense through a text-to-video generation task. This benchmark consists of 160 carefully crafted prompts designed to probe 27 distinct physical laws across four fundamental domains: mechanics, optics, thermal dynamics, and material properties. Unlike reference-based metrics, PhyGenBench employs PhyGenEval, a hierarchical evaluation framework that leverages advanced Vision-Language Models (VLMs) and Large Language Models (LLMs) to automate the assessment process. The evaluation is conducted in three progressive stages: Key Physical Phenomena Detection, which verifies the presence of specific physical events; Physics Order Verification, which checks the causal and temporal sequence of these events; and Overall Naturalness Evaluation, which rates the global realism of the video. This multi-stage approach ensures that the generated content is evaluated not just for visual quality, but for its adherence to fundamental physical principles.

Since PhyGenBench is originally designed for text-to-video (T2V) tasks, adapting it to our framework required input images to serve as initial conditions. To address this, we synthesized reference images using two distinct text-to-image (T2I) models: the 'gemini-2.5-flash-image-preview' (Comanici et al., 2025) and 'FLUX-schnell' (Esser et al., 2024) as shown in Fig. 10. In generating these inputs, we modified the original PhyGenBench prompts to explicitly depict the scene immediately preceding the onset of motion (*e.g.,* Original Prompt + ", static scene immediately before the action."). This temporal positioning facilitates a more precise comparison of the generated motion's intensity and velocity. In the main paper, we provide a comparative analysis of results based on both image sources to test generalizability. Furthermore, we acknowledge an inevitable domain gap between these synthetic starting frames and real-world imagery; thus, the evaluations reflect performance within a fully synthetic generation pipeline.

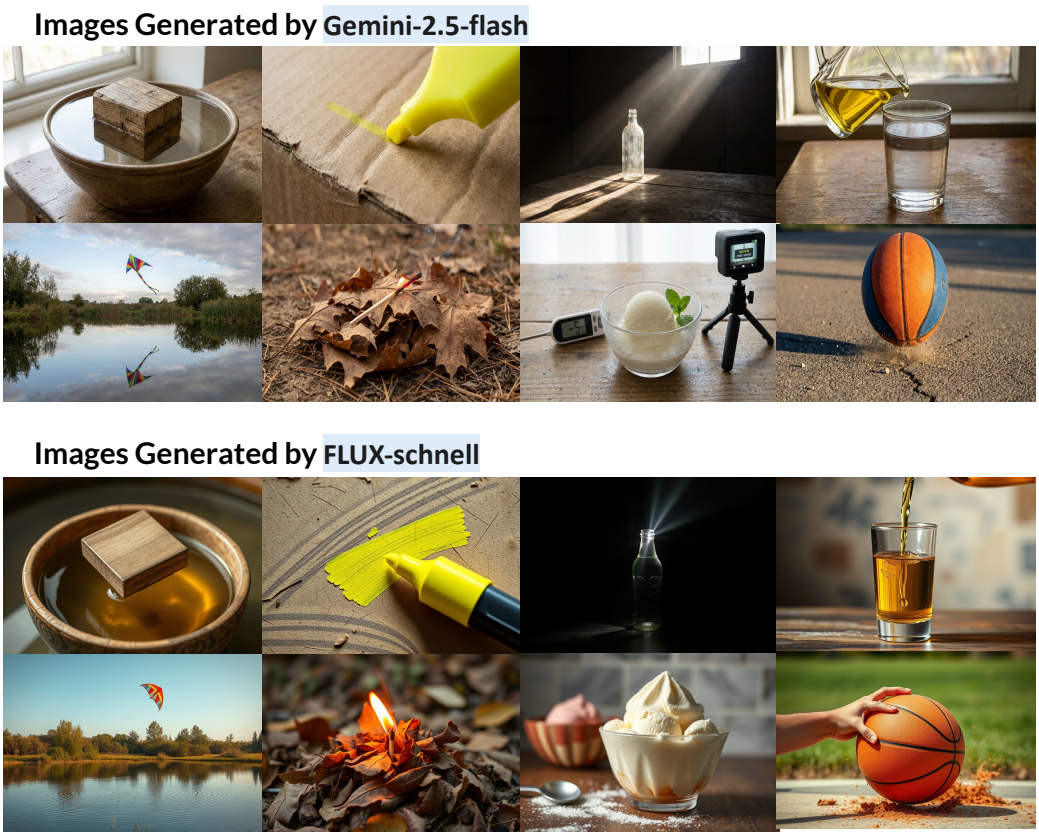


Figure 10. **Sample images from PhyGenBench generated by Gemini-2.5 Flash and FLUX-schnell, serving as input frames for the Image-to-Video (I2V) task.**

Table 8. **Human Evaluation Results.** We compare the Win Rate and Accuracy of our method against the baseline (CogVideoX-5B) across three categories. Note that Accuracy accounts for ties with a weight of 0.5.

	Win Rate (%)		Accuracy (%)	
Category	**Ours**	**Baseline**	**Ours**	**Baseline**
Physics Plausibility	**72.9**	27.3	**58.0**	42.0
Visual Quality	**77.5**	22.5	**60.8**	39.2
Prompt Alignment	**62.6**	37.4	**54.1**	45.9

Table 9. **Human Evaluation Results.** We compare the Win Rate and Accuracy of our method against the baseline (WAN 2.1) across three categories. Note that Accuracy accounts for ties with a weight of 0.5.

	Win Rate (%)		Accuracy (%)	
Category	**Ours**	**Baseline**	**Ours**	**Baseline**
Physics Plausibility	**81.9**	18.1	**67.1**	32.9
Visual Quality	**88.9**	11.1	**72.1**	27.9
Prompt Alignment	**78.4**	21.9	**63.1**	36.9

VBench Evaluation. Following the comprehensive evaluation protocol of (Yuan et al., 2026), we employed Physics-IQ and PhyGenBench to assess physical fidelity, alongside VBench (Huang et al., 2024) for general video quality. While our primary focus is on physical consistency, it is essential to ensure that the proposed method maintains high standards in fundamental generation metrics. Therefore, we utilized VBench to specifically evaluate key dimensions including Image Quality, Aesthetic Quality, Motion Smoothness, and Temporal Consistency. This verification ensures that the improvements in physical alignment do not compromise the overall visual fidelity and temporal coherence of the videos.

Human Study Protocol. To complement our quantitative metrics, we conducted a human preference study strictly following the protocol of (Yuan et al., 2026). We presented annotators with pairwise video comparisons (Ours vs. Baseline) in a randomized, blind manner to eliminate potential bias.The evaluation was conducted across three distinct criteria:

- *Physical Consistency:* Which video better adheres to real-world physical laws?
- *Text Alignment:* Which video better corresponds to the input text prompt?
- *Visual Quality:* Which video exhibits higher visual fidelity?

For each criterion, annotators were asked to select the superior video or indicate a Tie if both were of similar quality. The Win Rate was calculated as:

$$\text{Win Rate} = \frac{N_{\text{win}} + 0.5 \times N_{\text{tie}}}{N_{\text{total}}},$$

where N_{win}, N_{tie}, and N_{total} denote the number of wins, ties, and total comparisons, respectively.

To include neutral results in the evaluation rather than discarding them, we assign a score of 0.5 to each tie. The accuracy is calculated as follows:

$$\text{Accuracy} = \frac{N_{\text{win}} + 0.5 \times N_{\text{neutral}}}{N_{\text{total}}}$$

The overall results are shown below in Table 8.

F. Additional Experiement Results

F.1. Additional Ablation Studies

In this section, we present additional ablation studies that were not included in the main paper due to space constraints. Specifically, we evaluate the effectiveness of our adaptive scheduling by comparing it with alternative scheduling strategies.

We also analyze different guidance formulations to validate our choice of using Latent Delta guidance. Finally, we investigate the impact of the timestep range when applying this guidance. All experiments are conducted on CogVideoX-5B using the Physics-IQ benchmark.

F.1.1. Scheduling Ablations

We compare our linear decay schedule against four alternative scheduling strategies, all with matched total guidance magnitude (same $\lambda_0 = 0.05$ and active range). Table 10 presents the results.

Table 10. **Ablation on guidance scheduling strategies.** Linear decay achieves the best balance between early-stage motion anchoring and late-stage refinement flexibility.

Schedule	Formulation	Physics-IQ
Exponential	$\lambda_0 \cdot e^{-\alpha k}$	34.9
Cosine	$\lambda_0 \cdot \frac{1+\cos(\pi\rho)}{2}$	32.9
Constant	λ_0	31.3
Step	$\lambda_0 \cdot \mathbf{1}_{[k<k_{\text{end}}]}$	31.3
Linear (Ours)	$\lambda_0 \cdot \left(1 - \frac{k-k_{\text{start}}}{k_{\text{end}}-k_{\text{start}}}\right)$	**36.0**

The results reveal that gradual decay is essential for optimal performance. Constant scheduling (31.32) applies uniform guidance throughout, which over-constrains the model during later timesteps when high-frequency visual details should be refined freely. Step scheduling shows identical performance, confirming that abrupt termination provides no benefit over constant guidance within the active range.

Exponential decay (34.91) improves over constant by reducing guidance more rapidly, but its aggressive early decay ($e^{-\alpha k}$ drops to $< 10\%$ of λ_0 by mid-range) releases the motion constraint too quickly, allowing phase drift to accumulate. Cosine scheduling (32.92) suffers from the opposite problem: it maintains near-maximum guidance for too long before rapidly dropping, creating a discontinuity in the guidance profile.

Our linear decay achieves the best performance (36.0) by providing strong guidance during early timesteps when coarse motion structure is established (where phase preservation is most critical per Sec. 3), while smoothly releasing the constraint to allow the model to refine visual details. This matches the intuition from our phase erosion analysis: motion dynamics encoded in low-frequency phase are most vulnerable during early-to-mid denoising, and the guidance should taper proportionally as the generation progresses toward high-frequency refinement.

F.1.2. Guidance Formulation Ablation

We compare our Latent Delta guidance against three alternative formulations for transferring motion information from the prior. Table 11 shows the results.

Table 11. **Ablation on guidance formulations.** Latent Delta guidance on frame differences significantly outperforms alternatives.

Formulation	Definition	Physics-IQ
Normalized	$\frac{\mathbf{M}^{\text{prior}}-\mathbf{M}^{(k)}}{\|\mathbf{M}^{\text{prior}}\|_2}$	31.3
Direct Latent	$\mathbf{z}^{\text{prior}} - \mathbf{z}^{(k)}$	16.0
Second Order	$\mathbf{A}^{\text{prior}} - \mathbf{A}^{(k)}$	11.9
Latent Delta (Ours)	$\mathbf{M}^{\text{prior}} - \mathbf{M}^{(k)}$	**36.0**

- **Direct Latent** guidance ($\mathbf{z}^{\text{prior}} - \mathbf{z}^{(k)}$) directly constrains the latent toward the 2-step output. This dramatically fails (15.97) because it conflates static content with motion dynamics, the guidance fights against legitimate visual refinement that improves appearance while inadvertently degrading physics. This confirms that *what* is transferred matters as much as *how*.

- **Second Order** guidance uses frame accelerations $\mathbf{A} = \mathbf{M}_{2:F} - \mathbf{M}_{1:F-1}$ (second temporal derivative), motivated by the intuition that accelerations directly encode physical forces. However, this performs worst (11.89) because second-order differences amplify high-frequency noise present in the 2-step prior, which has lower visual fidelity. The guidance becomes dominated by noise rather than meaningful dynamics.

- **Normalized** guidance scales the delta by prior magnitude, intended to provide scale-invariant steering. This underperforms (31.34) because normalization disrupts the natural relationship between guidance magnitude and motion amplitude, large motions require proportionally larger corrections, which normalization prevents.

Our Latent Delta guidance operates on first-order frame differences without normalization, directly targeting the velocity field that encodes motion dynamics. Per our theoretical analysis (Sec. 4), frame deltas $\mathbf{M} = \mathbf{z}_{2:F} - \mathbf{z}_{1:F-1}$ encode motion phase information while being invariant to static scene content, achieving precise motion transfer without interfering with appearance refinement.

F.1.3. Hyperparameter Sensitivity.

We provide extended analysis of the three key hyperparameters introduced in Sec. 5: guidance strength λ_0, number of few inference steps (NFE), and guidance end step k_{end}. While the main paper presents the overall trends, here we discuss the underlying mechanisms and practical implications in detail.

Guidance Strength (λ_0). As shown in the main paper, performance exhibits a clear inverted-U pattern with peak at $\lambda_0 = 0.05$ (36.0). The mechanisms behind this are twofold. For weak guidance ($\lambda_0 = 0.03$, 34.1), the steering signal is insufficient to counteract the model's tendency toward phase drift during denoising, the latent trajectory deviates from the motion prior before correction can accumulate. For strong guidance ($\lambda_0 \geq 0.10$), the generation becomes over-constrained: the model cannot refine high-frequency visual details because guidance forces it to remain too close to the low-fidelity 2-step prior. At $\lambda_0 = 0.15$, performance degrades to 31.3, which is worse than the CogVideoX baseline (31.32). This demonstrates that guidance strength must balance two competing objectives: (1) anchoring the motion trajectory to preserve phase information, and (2) allowing sufficient freedom for the model to refine textures and details. The optimal $\lambda_0 = 0.05$ achieves this balance, providing enough correction to lock the phase evolution while permitting visual refinement orthogonal to motion.

few Inference Steps (NFE). Performance monotonically decreases as NFE increases: 2 steps achieves 36.0, while 10 steps degrades to 30.5. This result directly validates our core theoretical contribution, that phase erosion accumulates during denoising. Each additional step in the prior inference allows more phase corruption to occur, degrading the quality of the motion signal before it is even used for guidance. The steep drop from 2 to 5 steps (36.0 $\rightarrow$ 32.8) is particularly notable, confirming that phase corruption is most severe in early-to-mid denoising iterations where coarse motion structure is established. Beyond 5 steps, degradation continues but at a slower rate (32.8 $\rightarrow$ 30.5 from 5 to 10 steps), suggesting that the majority of phase damage occurs early. Practically, this means our method achieves optimal results with the fewest possible prior (2 steps), which also minimizes computational overhead.

Guidance End Step (k_{end}). Fixing $k_{\text{start}} = 0$, we vary k_{end} from 15 to 40 with $K_{\text{full}} = 50$. Performance peaks at $k_{\text{end}} = 25$ (36.0), corresponding to guidance during exactly the first half of denoising. The results reveal an interesting asymmetry: terminating guidance too early ($k_{\text{end}} = 15$, 35.8; $k_{\text{end}} = 20$, 35.0) causes mild degradation as motion structure is not yet fully established when guidance stops. However, extending guidance too long shows even milder effects ($k_{\text{end}} = 30$, 35.4; $k_{\text{end}} = 40$, 35.3), indicating that late-stage guidance neither helps nor significantly hurts, by this point, the coarse motion trajectory is already locked, and guidance becomes redundant rather than harmful. The relatively flat profile across $k_{\text{end}} \in [15, 40]$ (all scores within 35.0–36.0) demonstrates that our method is robust to this hyperparameter, simplifying practical deployment. We recommend $k_{\text{end}} = K_{\text{full}}/2$ as a principled default that balances motion anchoring with refinement freedom.

F.2. More qualitative results

In this section, we present further qualitative results in Fig. 16, Fig. 17, Fig. 18, and Fig. 19.

Note. Full-page figures are placed at the bottom of the document.

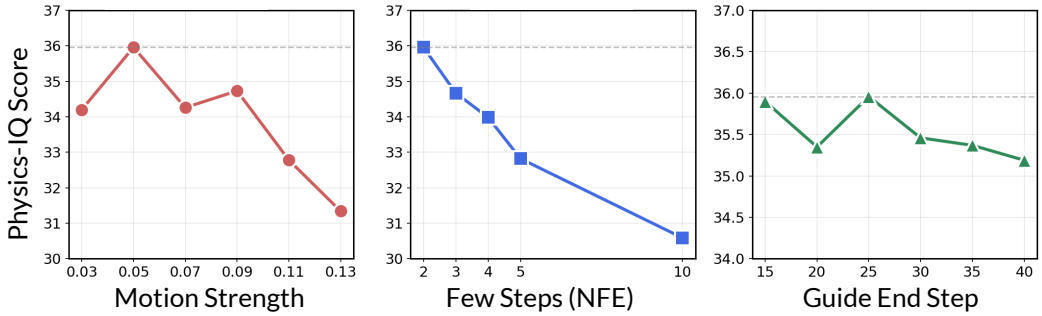


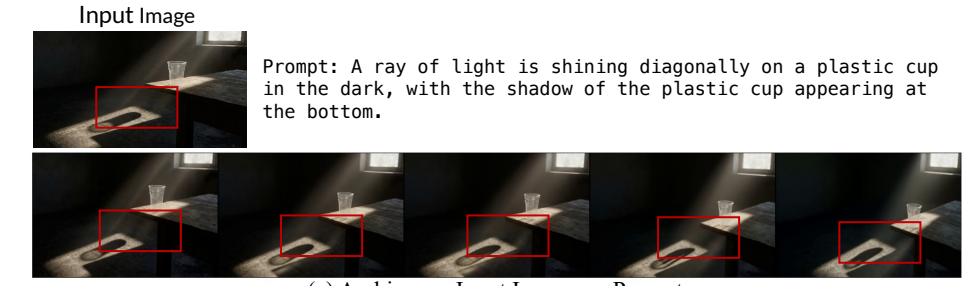
Figure 11. **Ablation Studies of PhaseLock**

G. Limitations and Further Discussions

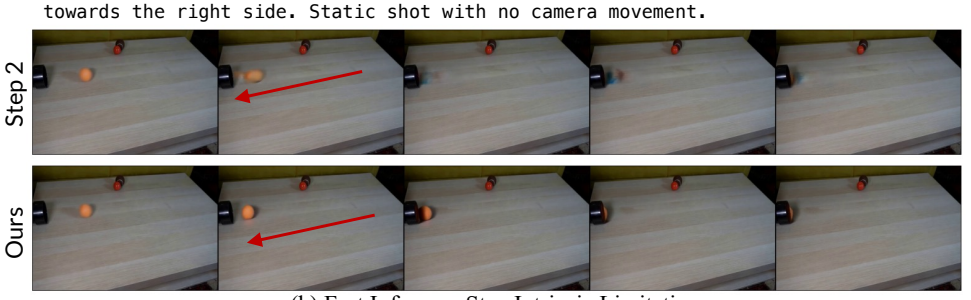
G.1. Limitations and Future Works

While PhaseLock demonstrates significant improvements in physical consistency, we acknowledge several limitations and failure cases inherent to our approach.

Dependence on few Inference Quality. Our method relies on the 2-step generation to provide a physically plausible motion prior. When the few inference itself produces incorrect physics, for instance, due to ambiguous input images or prompts that contradict physical intuition, the guidance will lock the high-fidelity generation to this erroneous trajectory (Fig. 12). Furthermore, due to the absence of further training, the system is liable to produce erroneous outputs even when the flawed Step 2 originates from the model's intrinsic limitations (Fig. 12). However, our guidance strength λ is moderate (not forcing exact replication), allowing the model some flexibility to deviate when strong textural cues conflict with the priors, preserving existing capability rather than creating new understanding.



(a) Ambiguous Input Images or Prompts



(b) Fast Inference Step Intrinsic Limitations

Figure 12. **Failure Cases of Our Method**

Architecture Specificity. PhaseLock is designed for diffusion-based video generation, where the iterative denoising process allows us to inject guidance at each step. The method is not directly applicable to Autoregressive models (e.g., VideoPoet (Kondratyuk et al., 2023), Emu Video (Girdhar et al., 2023)), and MAGI-1 (Teng et al., 2025). These generate frames sequentially without a denoising loop, so there is no opportunity to inject inter-frame guidance.

Trade-off with Creative Generation. By locking motion to the few inference prior, PHASELOCK may constrain artistic or physically unrealistic motions that users intentionally desire. This can be addressed by adjusting λ_0 or disabling guidance entirely for such use cases.

G.2. Further Discussions

In this section, we discuss the broader implications of our findings and outline potential extensions of PhaseLock.

G.2.1. WHY DOES PHASE EROSION OCCUR?

Our analysis identifies *what* happens (phase degrades fewer than magnitude) but does not fully explain *why* the denoising process exhibits this asymmetry. We hypothesize several contributing factors:

Training Objective Bias. Diffusion models are trained with MSE loss in pixel/latent space, which decomposes into magnitude and phase components in the frequency domain. By Parseval's theorem:

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathcal{F}(\mathbf{x}) - \mathcal{F}(\hat{\mathbf{x}})\|_2^2. \tag{34}$$

However, the gradient contribution from phase errors is modulated by magnitude:

$$\frac{\partial \text{MSE}}{\partial \phi} \propto A \cdot \sin(\phi - \hat{\phi}). \tag{35}$$

In high-frequency regions where A is small (fine details), phase gradients vanish, causing the model to prioritize magnitude alignment over phase alignment in these regions. Since motion dynamics often involve subtle positional shifts (small A, significant $\Delta\phi$), the training objective may inherently under-weight physical consistency.

Perceptual Loss Landscape. Human perception is more sensitive to texture (magnitude) than to small positional errors (phase). Training data curation and loss functions optimized for perceptual quality may implicitly encourage magnitude preservation at the expense of phase fidelity.

Coarse-to-Fine Generation Dynamics. Diffusion models generate global structure before local details (Choi et al., 2022). Phase encodes structure ("where things are"), so it is established early and becomes a target for subsequent refinement. If the model's capacity is allocated toward high-frequency detail generation in later steps, phase information may be treated as a "fixed" quantity to be overwritten rather than preserved.

Future work: A rigorous theoretical analysis of phase dynamics under different training objectives (MSE, perceptual loss, adversarial loss) could inform the design of physics-aware training procedures.

G.2.2. TOWARD PHYSICS-AWARE SAMPLERS

Our finding that phase erosion accumulates across denoising steps suggests that the **sampling algorithm itself** could be modified to minimize phase corruption.

Phase-Preserving Noise Schedules. Standard noise schedules (linear, cosine, EDM (Karras et al., 2022)) are designed to balance signal-to-noise ratio for visual quality. A physics-aware schedule could:

- Allocate more denoising capacity to early steps where phase is being established.
- Reduce the number of steps in the mid-to-late range where phase erosion is most severe.
- Introduce phase-weighted loss terms during sampling (though this would require gradient computation).

Frequency-Adaptive Sampling. Rather than applying uniform denoising across all frequency bands, an adaptive sampler could:

- Denoise low frequencies (magnitude-dominant) with standard schedules.

- Apply reduced noise perturbation to phase-dominant components.
- Implement band-specific step counts based on convergence.

This would require decomposing the latent into frequency bands at each step, which introduces computational overhead but could yield physics improvements without external guidance.

G.2.3. Extension to Other Modalities

While we focus on video generation, the phase erosion phenomenon may generalize to other sequential generation tasks.

(1) Audio Generation: In audio diffusion models (e.g., AudioLDM (Liu et al., 2023)), phase encodes temporal alignment and pitch, while magnitude encodes timbre and loudness. If similar phase erosion occurs, it could manifest as Temporal misalignment (sounds occurring at wrong times), Pitch drift, and Loss of rhythmic structure. A "PhaseLock for Audio" could extract rhythm/pitch priors from few inference and lock subsequent generation.

(2) 3D Generation: Extracting structural priors from few 3D inference could improve geometric plausibility.

(3) Multimodal Generation. For joint text-image-video generation, phase erosion could cause cross-modal misalignment (*e.g.,* narration describing action that doesn't match the visual motion). Locking temporal dynamics across modalities could improve coherence.

G.2.4. Theoretical Extensions

Information-Theoretic Analysis. Our empirical observation that phase degrades $\sim$18% from step 2 to step 50 while magnitude remains stable suggests an information-theoretic asymmetry. Formalizing this via Mutual information between phase/magnitude and physical consistency, Rate-distortion analysis of phase vs. magnitude under denoising, and Information bottleneck perspective on what is preserved vs. lost could provide theoretical grounding for our empirical findings.

Optimal Transport Perspective. Diffusion can be viewed as optimal transport from noise to data distribution. Phase erosion suggests the transport path is "curved" in a way that preserves magnitude but corrupts phase. Analyzing the Wasserstein geometry of this transport could reveal why phase is more vulnerable and how to design straighter (phase-preserving) paths.

Connection to Memorization vs. Generalization. Phase encodes specific structural details ("where exactly is the ball?"), while magnitude encodes general appearance ("what does a ball look like?"). The phase erosion phenomenon may relate to the model's tendency to generalize appearance while forgetting specific configurations, a form of "structural forgetting" during generation. This connects to broader questions about what diffusion models memorize vs. generalize.

G.2.5. Practical Extensions

User-Controllable Physics. Currently, PhaseLock extracts the motion prior automatically from few inference. Extensions could allow users to specify desired motion trajectories (e.g., "ball falls at 45° angle"), interpolate between multiple motion priors for controllable dynamics, and apply physics constraints from external simulators as soft priors.

Long-Video Generation. For videos longer than the model's native context, PhaseLock could be extended to extract motion priors for each temporal chunk, ensure consistency across chunk boundaries via overlapping guidance, and implement hierarchical priors at multiple temporal scales.

Real-Time Applications. The $1.06\times$ overhead of PhaseLock is already low, but for real-time applications (interactive generation, streaming), further optimization could include caching the 2-step prior across frames for video continuation, amortizing prior extraction across multiple outputs, and distilling the guidance into the model weights (one-time training cost).

G.2.6. SOCIETAL CONSIDERATIONS

Beneficial Applications. Physically consistent video generation enables:

- **Robotics**: Training policies on realistic simulated environments
- **Scientific visualization**: Accurate depiction of physical phenomena for education
- **Autonomous vehicles**: Generating diverse, physically plausible driving scenarios for testing
- **Accessibility**: Creating realistic visual descriptions for visually impaired users

Potential Misuse. Improved physical realism could make synthetic media harder to distinguish from real footage, potentially enabling:

- More convincing deepfakes
- Fabricated evidence of events that didn't occur
- Misinformation that exploits physical plausibility as a trust signal

Mitigation: We advocate for:

- Developing detection methods that identify PhaseLock-specific artifacts
- Watermarking generated content at the latent level
- Responsible disclosure practices that balance openness with harm reduction

Our contribution to understanding *how* physical consistency is achieved (phase preservation) also contributes to understanding *how* to detect synthetic content (by analyzing phase characteristics).

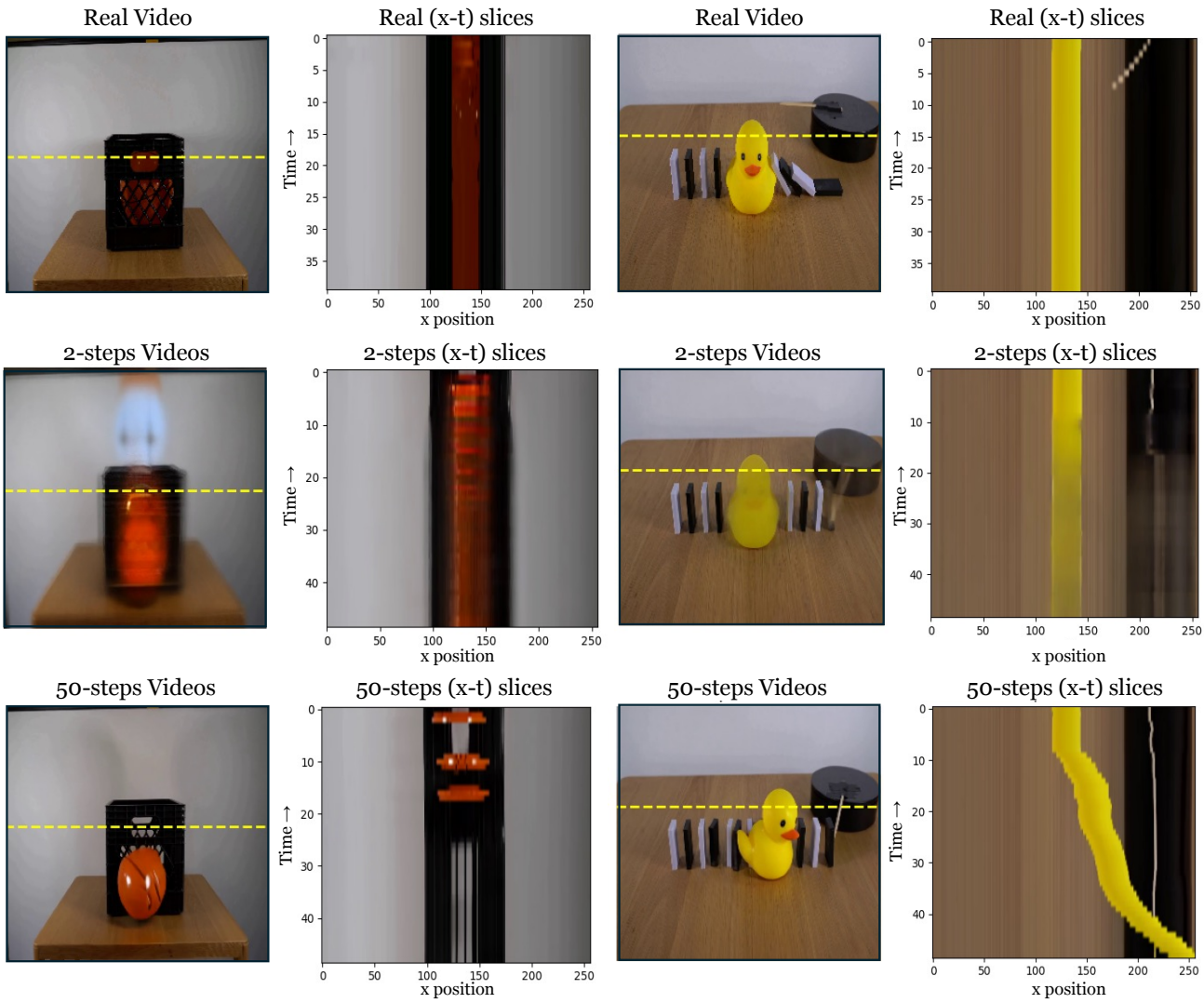


Figure 13. **Visualization of motion trajectories using spatio-temporal (x-t) slices**

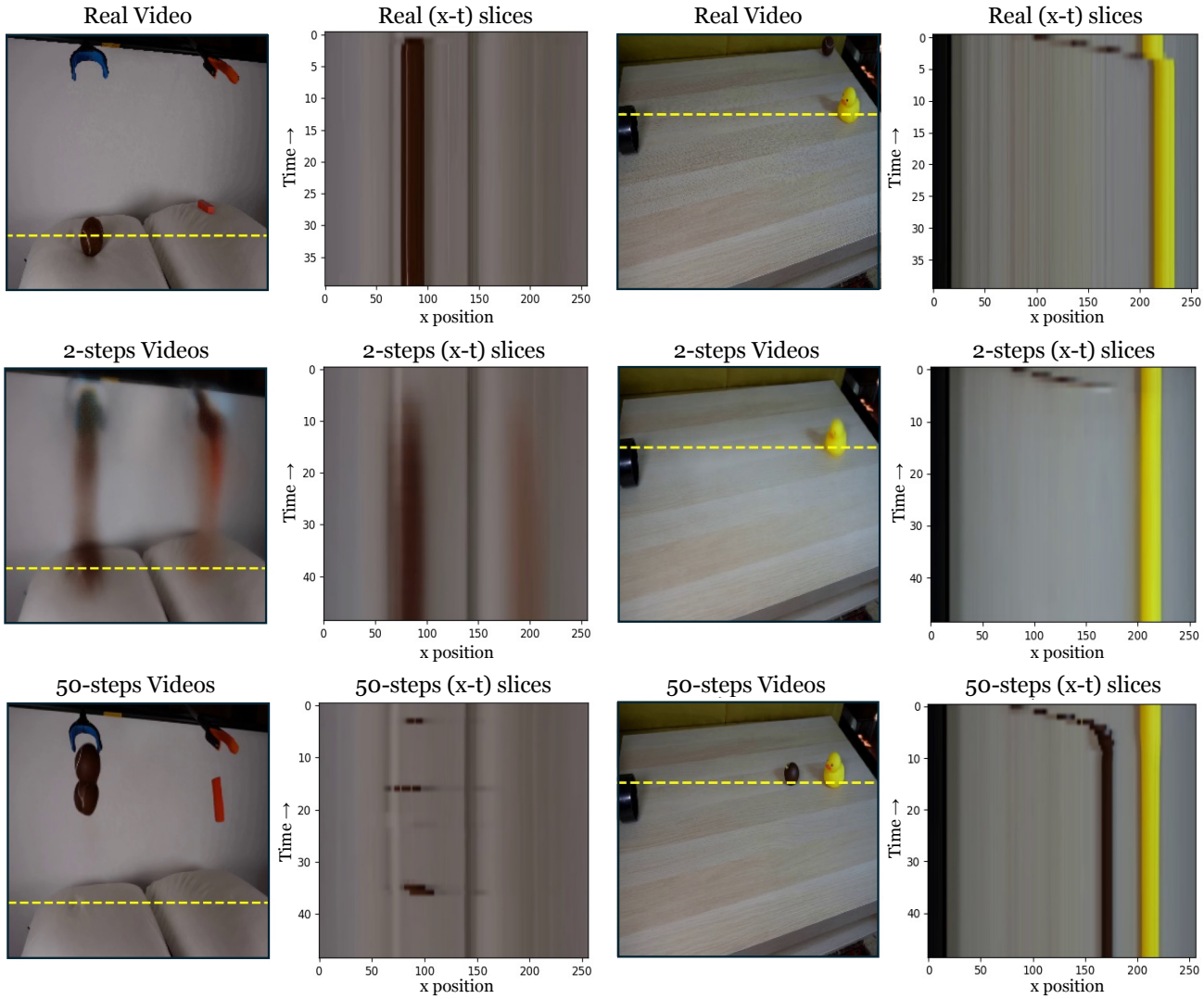


Figure 14. **Visualization of motion trajectories using spatio-temporal (x-t) slices**

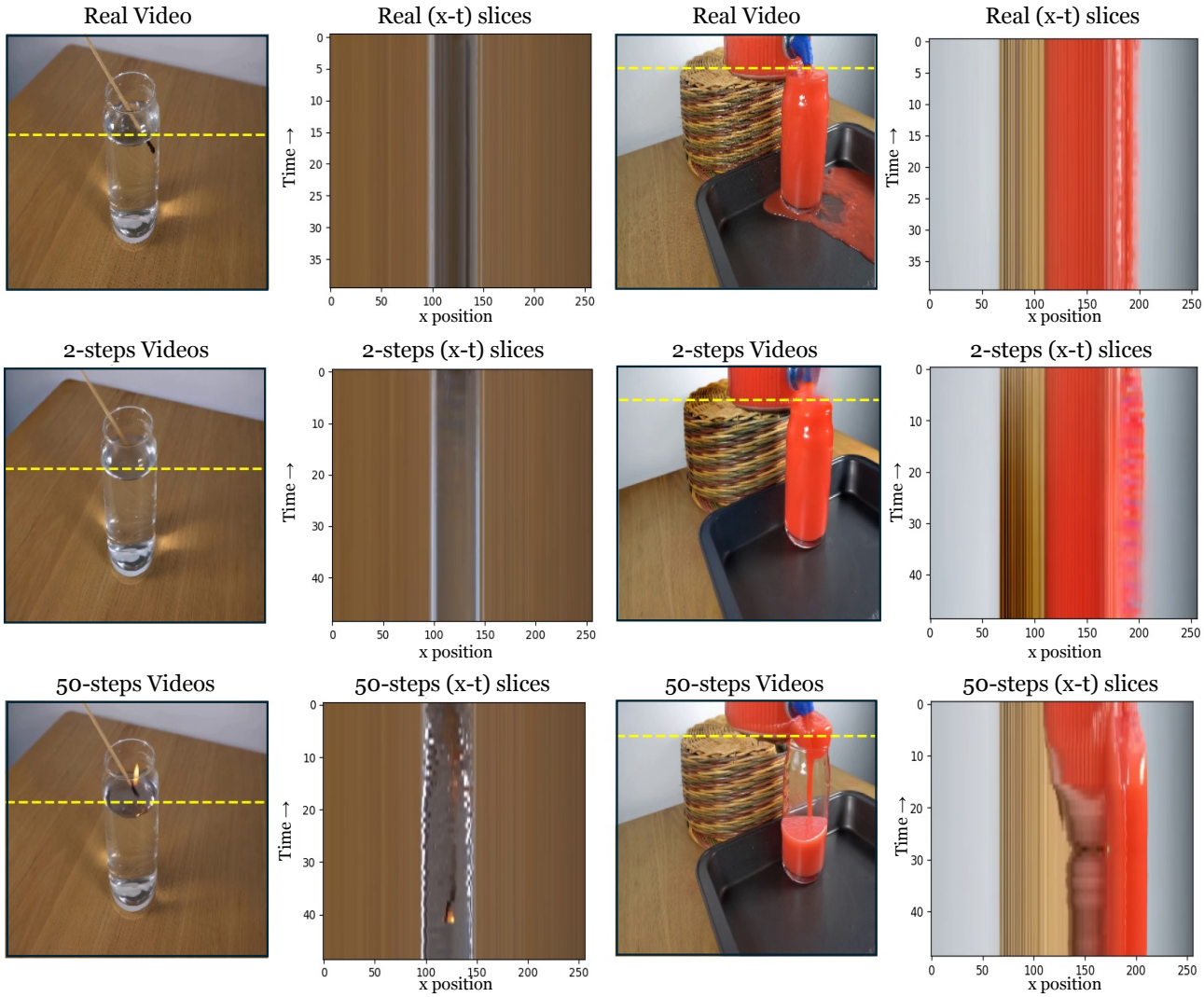


Figure 15. **Visualization of motion trajectories using spatio-temporal (x-t) slices**

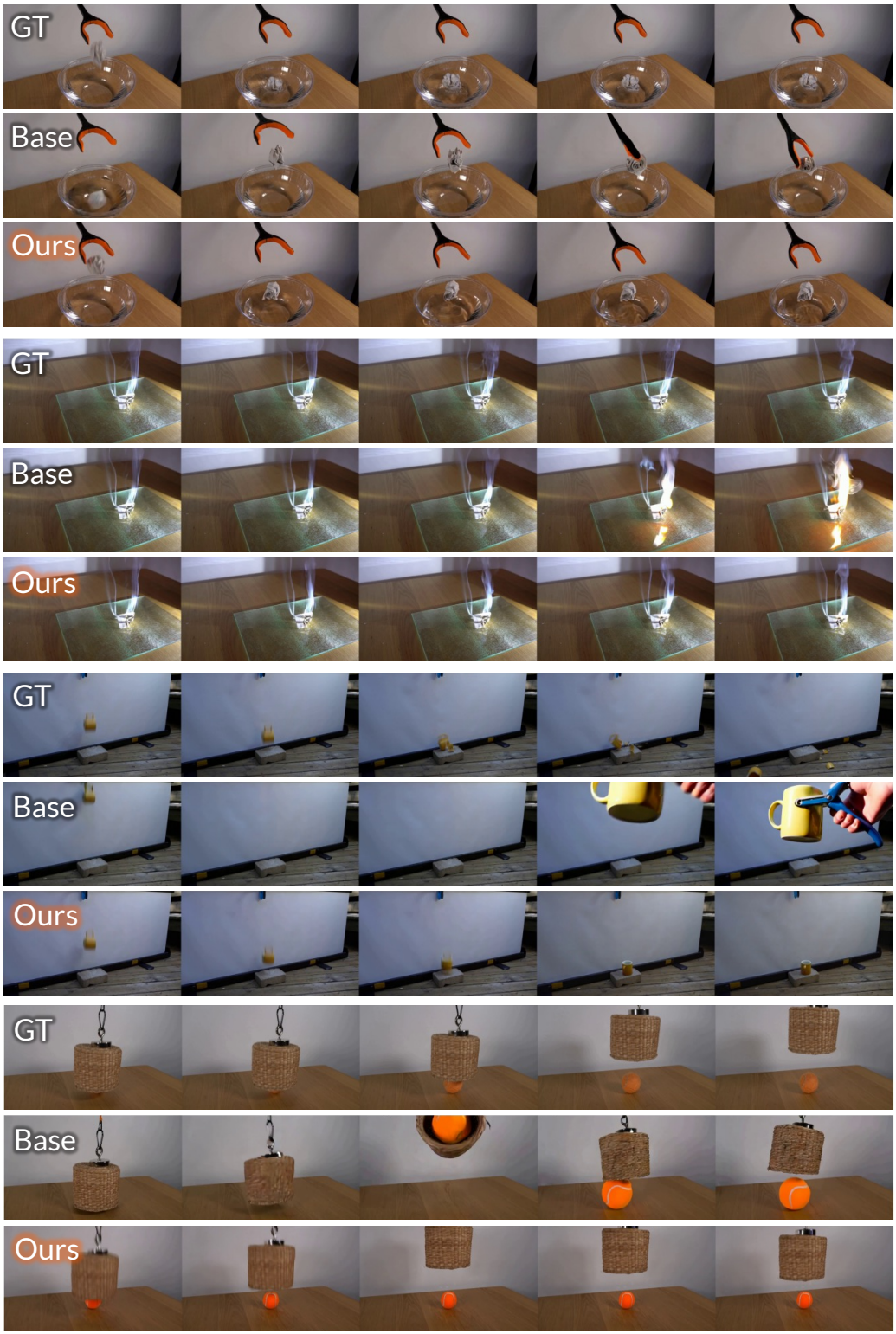


Figure 16. **Additional Qualitative Results for Physics-IQ Benchmark**

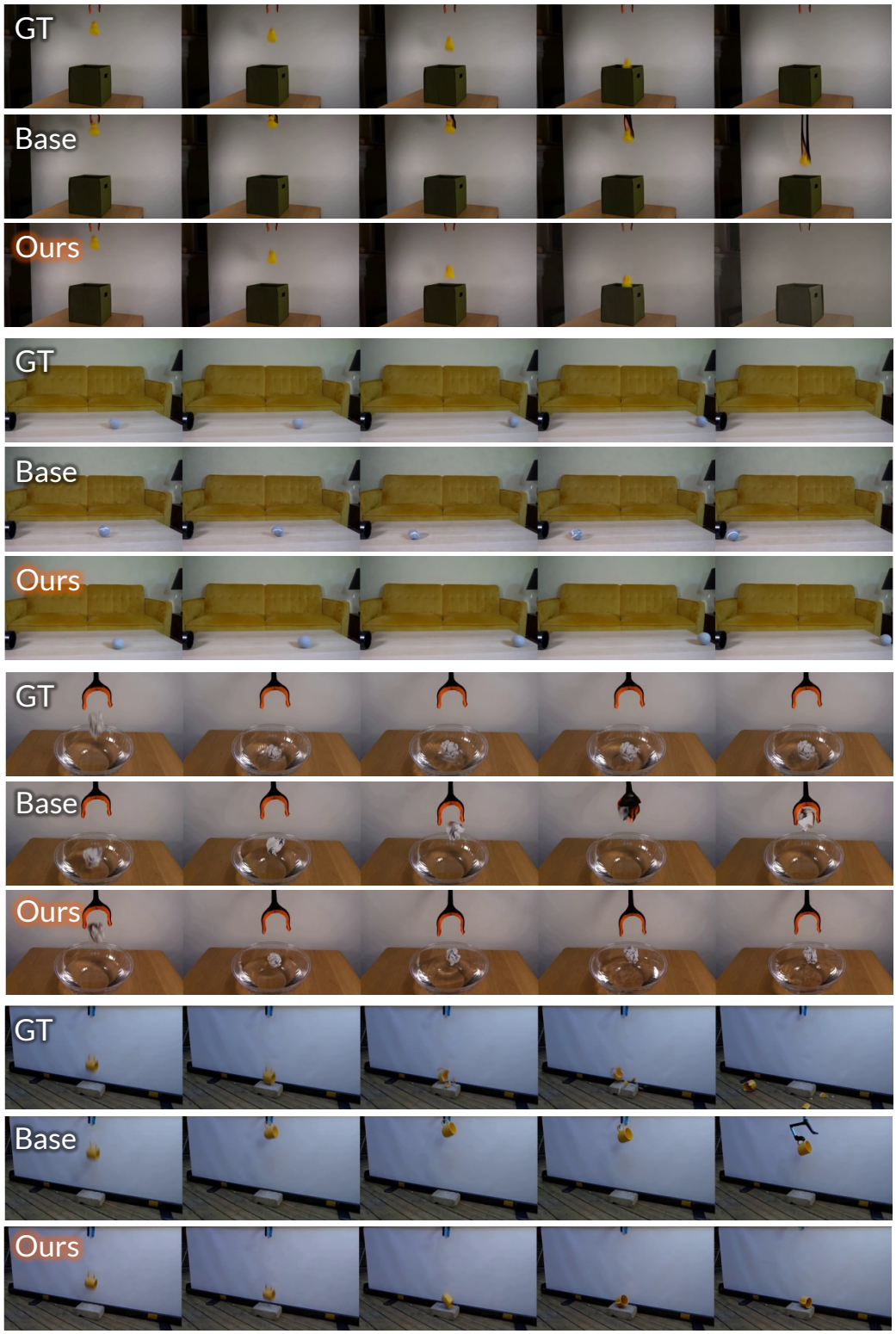


Figure 17. **Additional Qualitative Results for Physics-IQ Benchmark**

Figure 18. **Additional Qualitative Results for Physics-IQ Benchmark**

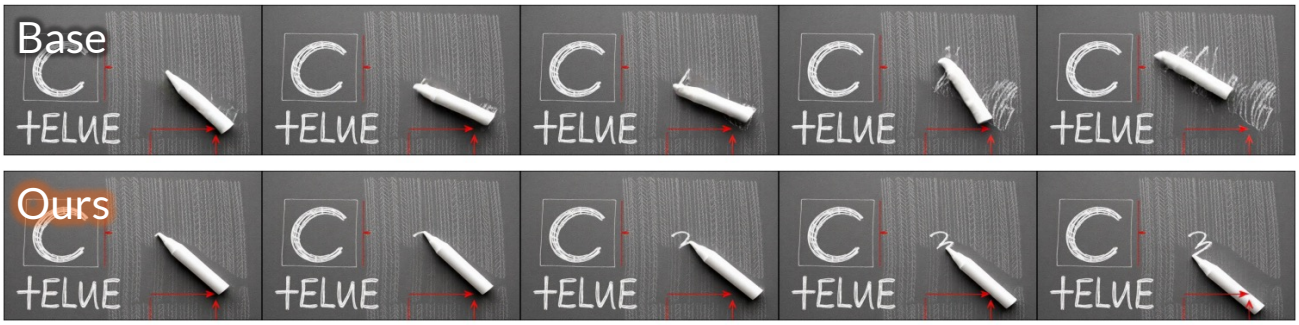
Figure 19. **Additional Qualitative Results for Physics-IQ Benchmark**

Figure 20. **Additional Qualitative Results for PhyGenBench**

Input Image

Input Text

A piece of white chalk is used to write on the rough, dark surface of a blackboard,showcasing the interaction between the chalk and the blackboard surface.



Input Image

Input Text

A clear plastic straw is slowly inserted into a glass of crystal-clear water, revealing the fascinating visual changes and reflections that occur as the straw interacts with the liquid.

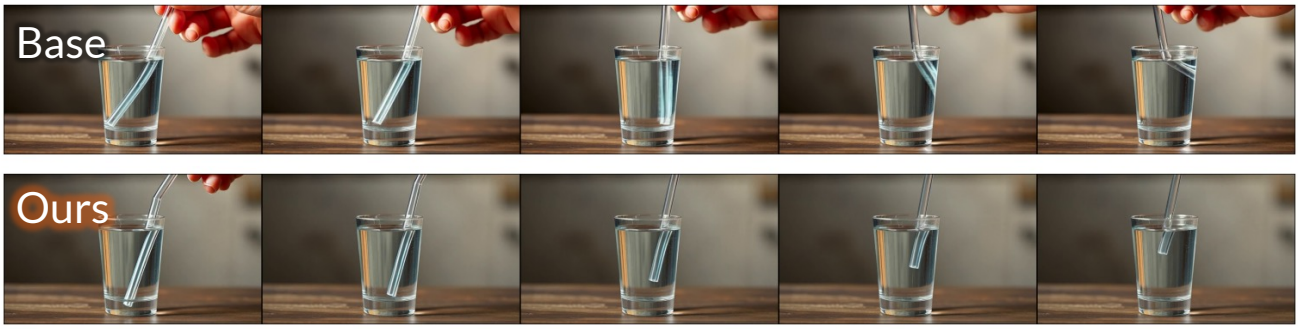


Figure 21. **Additional Qualitative Results for PhyGenBench**

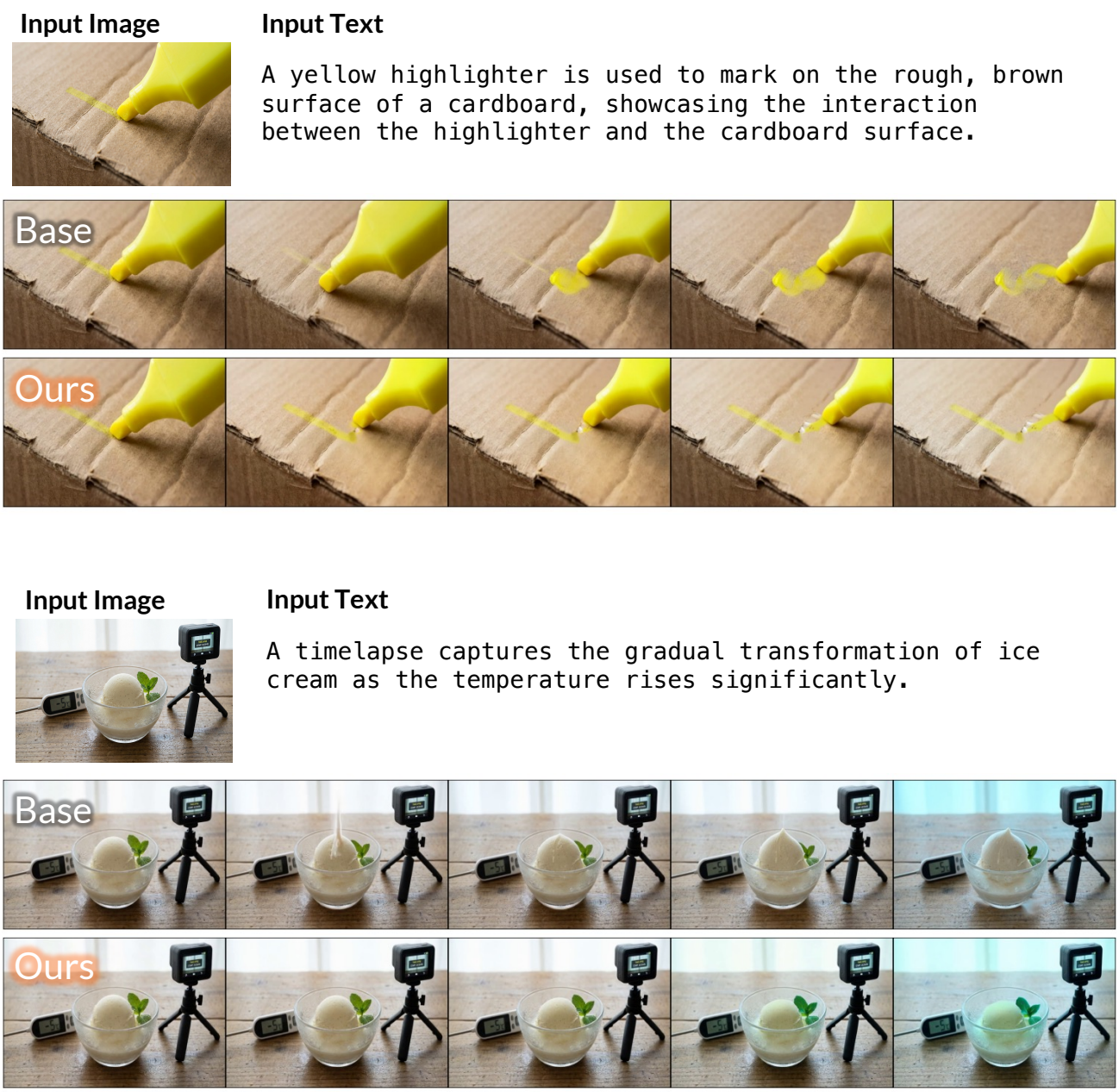


Figure 22. **Additional Qualitative Results for PhyGenBench**

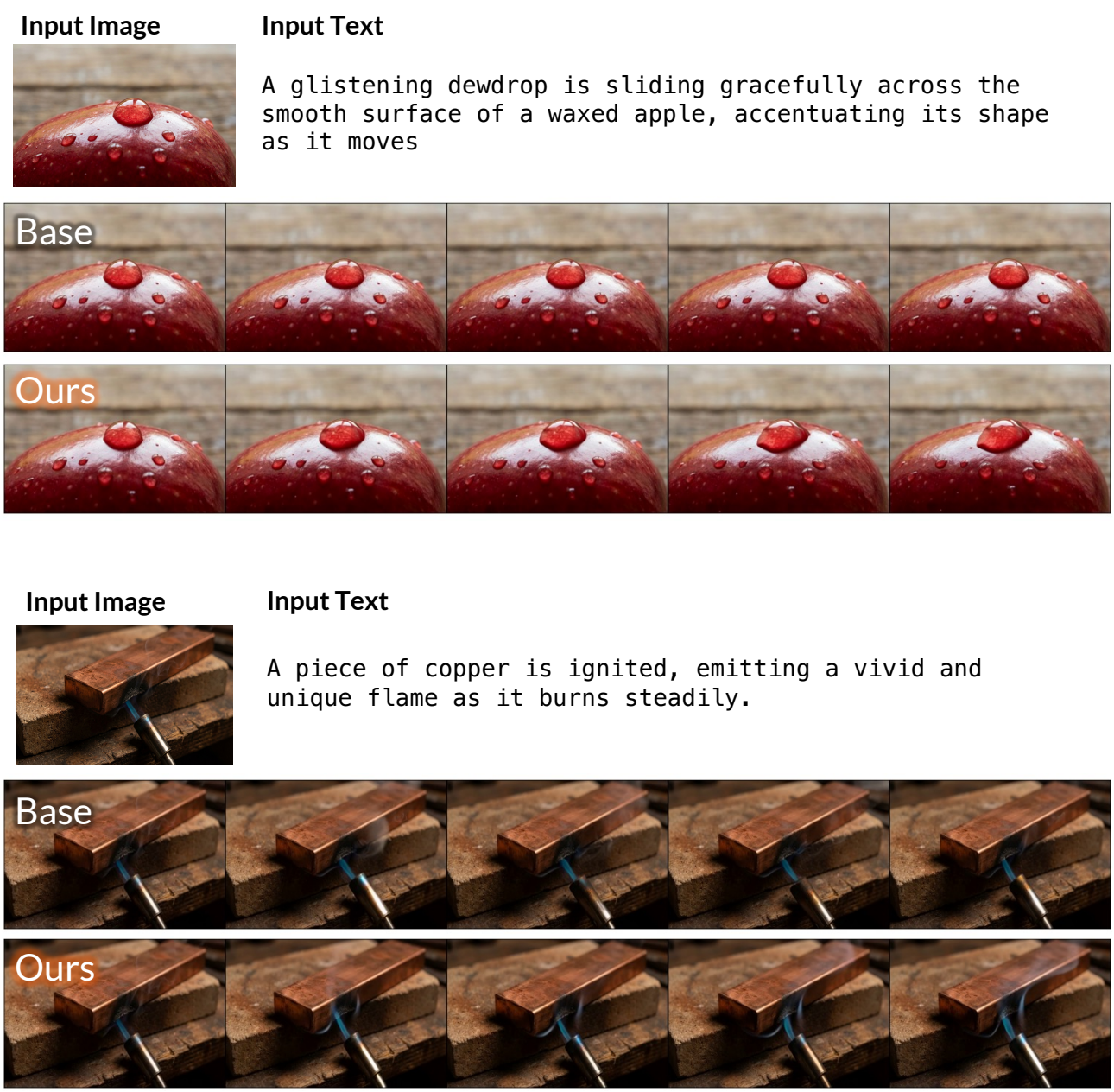


Figure 23. **Additional Qualitative Results for PhyGenBench**