
Physics-Guided Motion Loss for Video Generation Model

Bowen Xue¹ Giuseppe Claudio Guarnera² Shuang Zhao³ Zahra Montazeri¹

Abstract

Current video diffusion models generate visually compelling content but often struggle with physical motion, producing subtle artifacts like rubber-sheet deformations and inconsistent object motion. We introduce a frequency-domain physics prior that improves motion plausibility without modifying model architectures. Our method decomposes common motion patterns (translation, rotation, scaling) into lightweight spectral losses. Applied to Open-Sora, MVDIT, and Hunyuan, our approach improves both motion accuracy and action recognition by $\sim 11\%$ on average on OpenVID-1M (relative), while maintaining visual quality. Additional results on Wan 2.1-14B show consistent gains on video-quality and physics-oriented metrics. User studies show 74–83% preference for our physics-enhanced videos. It also reduces warping error by 22–37% (depending on the backbone) and improves temporal consistency scores. These results indicate that simple, global spectral cues are an effective drop-in regularizer for physically plausible motion in video diffusion.

1. Introduction

Diffusion-based video generation has recently achieved impressive frame quality. However, even flagship text to video diffusion systems still struggle with *physical motion*. Typical issues include rubber-like stretching, periodic flicker, and incorrect zooms or rotations. In short, frames can look good, but the motion often does not follow simple rules such as constant-velocity translation, rigid rotation, and uniform scaling.

Prior work to incorporate “physics” into video models mainly falls into four groups. (i) Flow/warping consistency

¹University of Manchester, Manchester, United Kingdom

²University of York, York, United Kingdom ³University of Illinois Urbana-Champaign, USA. Correspondence to: Bowen Xue <bowen.xue@manchester.ac.uk>.

helps reduce flicker but can break under large motion, occlusion, or brightness changes (Ho et al., 2020; Fleet & Jepson, 1990; Teed & Deng, 2020). (ii) Geometry-/3D-aware conditioning (e.g., depth or camera priors) improves rigidity but is domain-specific and computationally heavy in open settings (Harvey et al., 2022; Blattmann et al., 2023). (iii) Physics-inspired rules (e.g., constant-velocity penalties) help in narrow cases but do not scale well (Le Guen & Thome, 2020; Kataoka et al., 2020). (iv) Test-time refinements can smooth results but do not change the motion that the model actually learns.

Our idea is to regularize motion in the frequency domain, where basic physical motions leave simple, easy-to-detect patterns. Instead of matching pixels frame by frame, we look at global cues that summarize how energy moves over time. This view, formalized by a SIM(2) framework for translation, rotation, and scaling brings three practical benefits: it is global (not pairwise), more tolerant to brightness or small rendering errors, and helps separate translation, rotation, and scale without hand-tuned rules (Adelson & Bergen, 1985; Bracewell, 1956; Simoncelli & Heeger, 1998).

We turn these ideas into a frequency loss that encourages generated videos to show frequency patterns consistent with basic physical motion. A simple adaptive weighting focuses on whatever motion pattern is most supported in the current clip while remaining sensitive to mixtures. The loss is computed on a truncated spectrum for efficiency and drops into standard diffusion training without changing the backbone.

Contributions: (1) A frequency-based theoretical framework that connects basic physical motion to simple frequency-domain patterns, offering global, robust cues that reduce common motion ambiguities.

(2) A differentiable motion-aware regularizer with adaptive weighting that improves the learned motion behavior and handles mixed motion without hard classification.

(3) Experiments across multiple diffusion backbones show consistent gains in motion and temporal stability, visual quality and text–video alignment.

2. Related Work

Frequency-Domain Representation of Physical Motion.

The frequency representation of motion has deep roots: early

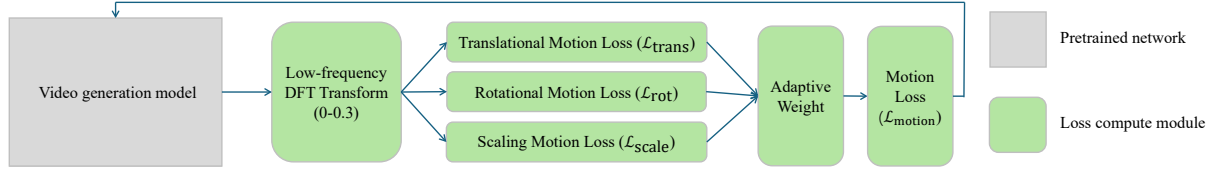


Figure 1. Our pipeline. From the generated video, we compute a low-pass 3D FFT, feed the spectral features to three motion losses (translation, rotation, scaling), and adaptively combine them into $\mathcal{L}_{\text{motion}}$ for training.

studies modeled motion perception in frequency space (Watson & Ahumada, 1985), and phase consistency was shown to encode motion, inspiring optical-flow methods (Fleet & Jepson, 1990). Translational motion was characterized spectrally (Simoncelli & Heeger, 1998); rotations yield annular energy with discrete temporal peaks (Bracewell, 1956). Frequency analysis supports motion energy filtering (Adelson & Bergen, 1985) and multiband decompositions for motion layering and segmentation (Wang & Adelson, 1994). Recent work exploits spectral traits for motion classification (Sevilla-Lara et al., 2016) and improves interpolation via frequency cues (Xue et al., 2019). For evaluation, spectral metrics quantify spatiotemporal coherence of generated videos (Tsfaldet et al., 2018), though most use frequency only at *evaluation* time rather than in training.

Physics-Constrained Video Generation. Integrating physics into generative modeling is emerging, including video prediction with physical consistency (Le Guen & Thome, 2020; Kataoka et al., 2020) and recent physics-aware video generation methods. PhysGen (Liu et al., 2024a) relies on scene-specific physics simulation together with perception and rendering modules such as segmentation, depth/normal estimation, inpainting, and rendering. Concurrent work such as PhyGDPO (Cai et al., 2026) explores preference optimization for improving physical plausibility in text-to-video generation. MotionCraft (Savant Aira et al., 2024) is a zero-shot latent-warping method driven by externally simulated optical flow, while our method is a training-time frequency-domain regularizer for video diffusion backbones.

Deep Learning-Based Video Generation Models. GAN-based methods separate static/dynamic components (Vondrick et al., 2016), decouple temporal and image generators (Saito et al., 2017), or disentangle motion/content (Tulyakov et al., 2018). Autoregressive models include flow-based VideoFlow (Kumar et al., 2020) and transformer tokenization (Weissenborn et al., 2020), but face temporal and computational limits. Latent-space approaches leverage VAEs and decomposed latents for high-resolution, temporally coherent generation (Villegas et al., 2019).

Diffusion Models for Video Generation. Diffusion has advanced video quality via joint spatiotemporal denoising (Ho

et al., 2020), spatiotemporal U-Nets (Harvey et al., 2022), and latent-space diffusion for efficiency (Blattmann et al., 2023). Conditional generation scales further with text and multimodal conditioning (Singer et al., 2022; Yin et al., 2023) and motion modules (Guo et al., 2024). Despite rapid progress, including Sora (Brooks et al., 2024) and Step-Video-T2V (Ma et al., 2025), maintaining physically plausible motion remains challenging.

3. Physics-Guided Motion-Aware Loss Function

State-of-the-art video diffusion models are typically trained with data-driven objectives plus optional optical-flow or temporal-smoothness terms (Ho et al., 2020; Harvey et al., 2022; Blattmann et al., 2023; Guo et al., 2024). While these reduce flicker, they do not explicitly encode basic physical motion (constant-velocity translation, rigid rotation, uniform scaling), so artifacts that *look* like motion—rubber-sheet deformations, periodic flicker, scale/zoom glitches—persist even in simple scenes (Brooks et al., 2024; Ma et al., 2025).

We ground our remedy in frequency space. Within a unified SIM(2) spectral framework (translations, rotations, uniform scalings) (Sharma & Duits, 2015)¹ spectral geometry, physically plausible motions exhibit simple slice-wise signatures: (i) *rotation*: energy aligns with tilted lines $\omega_t + m\Omega = 0$ in (m, ω_t) and concentrates annularly; (ii) *scaling*: radial-temporal gradients align and the radial spectral centroid shows a clear monotone trend; (iii) *translation*: energy lies near a plane in $(\omega_x, \omega_y, \omega_t)$.

Guided by these diagnostics, we add a small, differentiable frequency-domain regularizer on $\hat{x}_0$ composed of three *slice-consistent* losses (translation, rotation, scaling) with adaptive weighting, because full SIM(2) hyperplane suffers from (1) cross-slice interference and energy double counting (mixing $\{\omega_x, \omega_y\}$ with m, ν), (2) ill-conditioning under mixed motions when $|m|$ or $|\nu|$ excitation is weak, and (3) coupled weighting that obscures which failure mode is being

¹The 2D similarity group (translations, rotations, and uniform scalings): it acts on image-plane points as $x' = s R(\theta) x + t$ with $t \in \mathbb{R}^2$, $R(\theta) \in \text{SO}(2)$, and $s > 0$ (4 DoF). Shear and anisotropic scaling are excluded.

corrected.

3.1. Frequency-Domain Characteristics of Physical Motion

Different types of physical motion exhibit unique and distinguishable features in the frequency domain, providing a theoretical foundation for our loss function design. We first briefly outline these features and then discuss how we design loss functions based on them.

SIM(2) spectral manifold. Consider a short temporal window where the dominant motion is well-approximated by a similarity transform (Hartley & Zisserman, 2004) (translation $v = (v_x, v_y)$, in-plane rotation with angular velocity Ω , and isotropic scaling rate $\alpha = \dot{\sigma}$ in log-radius). Let $\hat{V}(\omega_x, \omega_y, \omega_t)$ be the FFT-based (Oppenheim et al., 1999) spatiotemporal spectrum. Passing to polar coordinates (ρ, θ) in the spatial frequency plane and expanding along the angular and log-radial axes yields harmonic indices $m \in \mathbb{Z}$ and $\nu \in \mathbb{Z}$. Then the ideal SIM(2) motion concentrates spectral energy ($E(\omega_x, \omega_y, \omega_t) = |\hat{V}(\omega_x, \omega_y, \omega_t)|^2$) on a single hyperplane in $(\omega_x, \omega_y, m, \nu, \omega_t)$:

$$\omega_t + v_x \omega_x + v_y \omega_y + \Omega m + \alpha \nu + b_0 = 0, \quad (1)$$

b_0 is a regression intercept that absorbs residual phase offsets and discretization bias; see App A.1 for the windowing convention. Three classical facts are recovered as special cases: (i) **translation**: $\omega_t + v_x \omega_x + v_y \omega_y = 0$ (a plane in $(\omega_x, \omega_y, \omega_t)$); (ii) **rotation**: $\omega_t + \Omega m = 0$ (tilted lines in (m, ω_t) , since the m -th angular harmonic acquires a temporal factor $e^{-im\Omega t}$); (iii) **scaling**: $\omega_t + \alpha \nu = 0$ (tilted lines in (ν, ω_t)). These equal the translation/rotation/scaling slices of the SIM(2) spectral hyperplane; derivations are in App. A.2, and practical choices (polar/log-radius resampling, energy/observability gating, robust regression) are in App. A.6. Figure 2 visualizes these spectral slices on synthetic SIM(2) sequences.

Multi-object and mixed motions. Although (1) is derived for a single SIM(2) motion in a short window, by linearity of the 3D FFT, multiple SIM(2) motions give rise to an approximately additive superposition of their spectral patterns. This behavior is visible in the bottom rows of Fig. 2, where two independently translating objects and a moving and shrinking object generate multiple structures in the translation, rotation, and scaling slices.

Energy-weighted unified residual. From the observed spectrum (the 3D FFT of the current $T \times H \times W$ video window; energy $E = |\hat{V}|^2$) we build samples (ϕ_i, b_i, w_i) with design vector

$$\phi_i = [\omega_x, \omega_y, m, \nu, 1], \quad b_i = -\omega_t,$$

and energy/observability weight w_i (details in App. A.1 energy gate, low- m/ν suppression, Huber/Charbonnier ro-

bustification). We estimate $\theta = [v_x, v_y, \Omega, \alpha, b_0]^\top$ by weighted ridge regression

$$\hat{\theta} = \arg \min_{\theta} \sum_i w_i (\phi_i \theta - b_i)^2 + \lambda \|\theta\|_2^2.$$

We then define the unified residual as $\mathcal{L}_{\text{uni}} = \frac{\sum_i w_i (\phi_i \hat{\theta} - b_i)^2}{\sum_i w_i}$. If the window follows a SIM(2) motion, $\mathcal{L}_{\text{uni}}$ tends to 0 as $T \rightarrow \infty$. For general motions and finite windows, $\mathcal{L}_{\text{uni}}$ upper-bounds and controls the off-hyperplane energy fraction under the band tolerance Δ (cf. Lemma B.2 and Lemma B.3). *Sketch.* Combine the classical translation plane, rotational angular-harmonic identity $\delta(\omega_t + \Omega m)$, and log-radial shift identity $\delta(\omega_t + \alpha \nu)$ under Parseval; see App A.2 and A.3 for details.

Differentiable WLS. We estimate motion parameters with an energy-weighted ridge least-squares fit (differentiable solve; $\lambda = 10^{-3}$, jitter ($\varepsilon = 10^{-8}$), FP32; pseudo-inverse fallback).

3.1.1. FREQUENCY-DOMAIN CHARACTERISTICS OF TRANSLATIONAL MOTION

For translational motion with constant velocity (v_x, v_y) , its spatiotemporal representation is:

$$V(x, y, t) = V_0(x - v_x t, y - v_y t). \quad (2)$$

In the frequency domain, this motion concentrates energy on an *affine* plane:

$$\omega_t + \omega_x v_x + \omega_y v_y + b_0 = 0, \quad (3)$$

where b_0 absorbs windowing/phase conventions (in the ideal unwrapped case $b_0=0$). This is the translation slice of the SIM(2) model (cf. Eq. (1)) (Adelson & Bergen, 1985; Bracewell, 1956; Simoncelli & Heeger, 1998). The normal vector $(v_x, v_y, 1)$ is directly related to the motion velocity. In our synthetic example, the panel in the first row, second column of Figure 2 shows that the energy in $E_{x,t}(\omega_x, \omega_t)$ (obtained by averaging over ω_y) concentrates along an approximately straight ridge, consistent with the affine plane in (3).

3.1.2. FREQUENCY-DOMAIN CHARACTERISTICS OF ROTATIONAL MOTION

Under in-plane rotation with angular velocity Ω ,

$$V(r, \theta, t) = V_0(r, \theta - \Omega t).$$

The spatiotemporal frequency analysis implies two signatures (see App. A.2): (i) *annular* spatial energy concentration (from the Bessel-type radial response of angular harmonics), and (ii) *tilted lines* in the (m, ω_t) plane given by

$$\omega_t + m\Omega = 0,$$

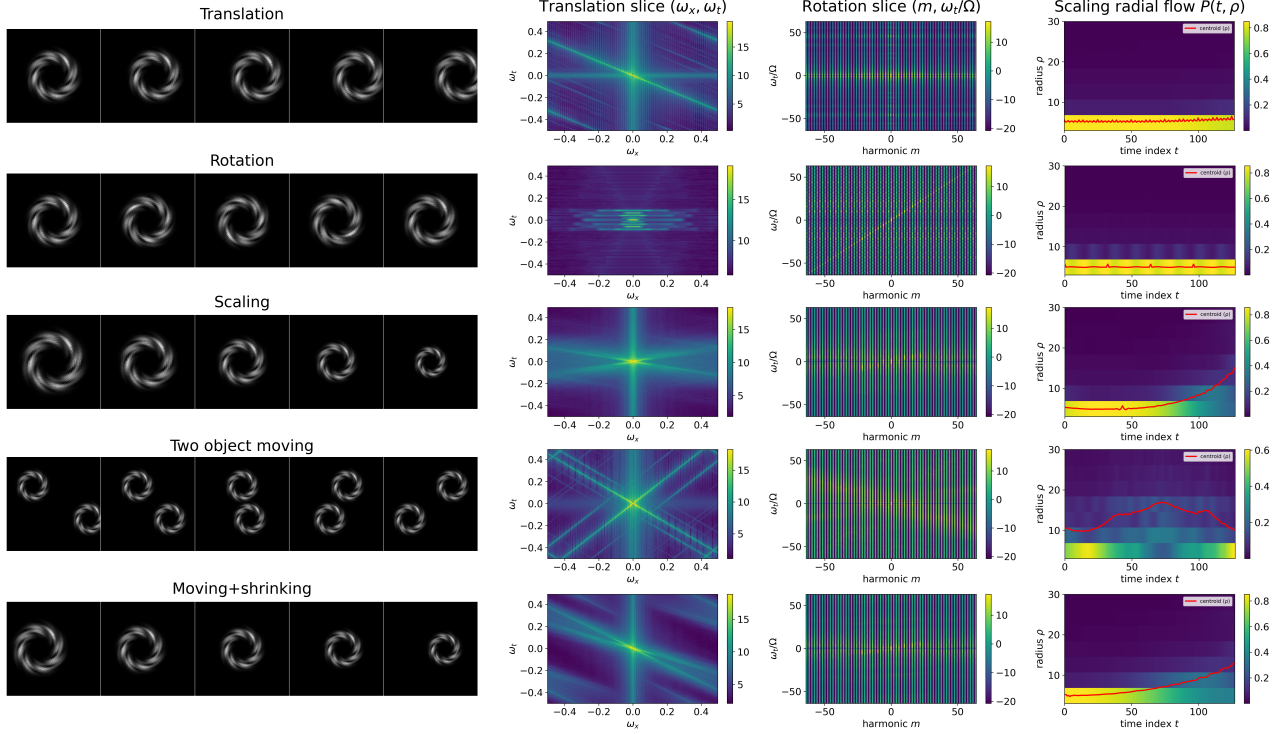


Figure 2. Visualization of SIM(2) spectral signatures on synthetic sequences. Rows show different motions; columns show sample frames, the translation slice, the rotation slice, and the scaling radial-flow map with radial centroid (red). Pure SIM(2) motions (top rows) produce the expected simple structures in each slice, while the multi-object and mixed-motion rows show superpositions that remain visually separable.

i.e., the m -th angular harmonic carries a temporal tone at $m\Omega$. Under finite temporal windows these appear as narrow bands (cf. Lemma B.3). This is precisely the rotational slice of the unified SIM(2) plane in Eq. (1) (Bracewell, 1956; Adelson & Bergen, 1985; Simoncelli & Heeger, 1998; Fleet & Jepson, 1990). In Figure 2 (third column in each row), we normalize temporal frequency by the ground-truth angular velocity Ω , so that the ideal relation $\omega_t + m\Omega = 0$ becomes straight rays $\omega_t/\Omega \approx -m$ passing through the origin in the $(m, \omega_t/\Omega)$ plane.

3.1.3. FREQUENCY-DOMAIN CHARACTERISTICS OF SCALING MOTION

A spatial scaling by a factor $s > 0$ obeys the Fourier-scaling law in 2D:

$$\mathcal{F}\{V_0(x/s, y/s)\}(\omega_x, \omega_y) = s^2 \hat{V}_0(s\omega_x, s\omega_y),$$

so scaling induces a radial reallocation of spectral energy (Brigham, 1988; Oppenheim et al., 1999; Stephane, 1999). For a time-varying scale $s(t) = e^{\sigma(t)}$, passing to log-polar coordinates $(\rho, \theta) \mapsto (\xi = \log \rho, \theta)$ turns scaling into a translation $\xi \mapsto \xi - \sigma(t)$; consequently, after angular/log-radial expansions the energy in (ν, ω_t) concentrates along

the line

$$\omega_t + \alpha \nu = 0, \quad \alpha = \dot{\sigma}(t),$$

which is precisely the scaling slice of our unified SIM(2) model (cf. Eq. (1)) (Lindeberg, 1993; Reddy & Chatterji, 1996). Operationally, this appears as a “radial energy flow”: zoom-in (increasing s) shifts energy to lower spatial frequencies, and zoom-out to higher ones. The rightmost column of Figure 2 shows the corresponding radial-flow map $P(t, \rho)$ and the radial centroid $\rho_c(t)$ (red); for the pure scaling row, the centroid follows a clear monotone trend, matching the expected radial energy flow induced.

For training-time measurements we adopt two simple proxies. (i) A radial-temporal gradient alignment score C_{flow} (dot product of normalized $\nabla_\rho E$ and $\nabla_t E$) captures the strength of radial-flow alignment. (ii) The temporal trend of the radial spectral centroid $\rho_c(t)$ provides a scale-rate proxy (derivations and bounds in App. §A.4–A.5).

3.1.4. COMPUTABLE BOUNDS FOR WINDOW/INTERPOLATION

For a chosen temporal window h (Hann), the RHS of Lemma B.3 gives a closed-form/lookup bound on $\varepsilon_{\text{win}}(\Delta)$ as a function of (T, Δ) . The effect of bilinear polar/log-

radius resampling can also be controlled under standard local smoothness assumptions on the spectrum.

3.2. Translational Motion Loss

For constant-velocity translation (v_x, v_y) , the spectral support concentrates near a plane in $(\omega_x, \omega_y, \omega_t)$ as in Eq.(3). We estimate the parameters via energy-weighted least squares (WLS). Let $A_i = (\omega_{x,i}, \omega_{y,i}, 1)$, $b_i = -\omega_{t,i}$, $\beta_{\text{tr}} = [v_x, v_y, b_0]^\top$. With weights $\mathbf{W}_{ii} \geq 0$ (energy/observability gating; App. A.6), the ridge-WLS estimator and the normalized residual are

$$\hat{\beta}_{\text{tr}} = \arg \min_{\beta_{\text{tr}}} \sum_i \mathbf{W}_{ii} (A_i \beta_{\text{tr}} - b_i)^2 + \lambda \|\beta_{\text{tr}}\|_2^2,$$

$$\mathcal{L}_{\text{trans}} = \frac{\sum_i \mathbf{W}_{ii} (A_i \hat{\beta}_{\text{tr}} - b_i)^2}{\sum_i \mathbf{W}_{ii}}.$$

This plane model is the translation slice of our SIM(2) framework and is stable under short temporal windows; non-constant velocities appear as bandwidth broadening along ω_t and are naturally penalized by the residual (see App. A.6 for windowing and tolerance).

3.3. Rotational Motion Loss

Rotational motion exhibits two complementary spectral signatures: (i) annular spatial concentration; (ii) energy alignment along tilted lines $\omega_t + \Omega m = 0$ in the (m, ω_t) plane. We therefore adopt the ring-concentration term and tilted-line energy ratio that matches the rotational slice of the unified SIM(2) model.

We estimate the angular velocity with an energy-weighted least squares:

$$\Omega^* = - \frac{\sum_\rho \sum_{m \neq 0} \sum_{\omega_t} |\tilde{C}_m(\rho, \omega_t)|^2 \omega_t m}{\sum_\rho \sum_{m \neq 0} \sum_{\omega_t} |\tilde{C}_m(\rho, \omega_t)|^2 m^2}.$$

The tilted-line energy ratio and ring concentration are

$$C_{\text{rot}} = \frac{E_{\text{line}}}{E_{\text{all}}}, \quad E_{\text{line}} = \sum_{\rho, m \neq 0, |\omega_t + m\Omega^*| \leq \Delta} |\tilde{C}_m|^2,$$

$$C_{\text{ring}} = 1 - \frac{H_{\text{ring}}}{\log N_r}.$$

$(\tilde{C}_m(\rho, \omega_t))$ denotes the time-DFT of the angular harmonic $C_m(\rho, t)$; H_{ring} is the entropy of energy over N_r concentric rings; Δ is one temporal-frequency bin in the condition $|\omega_t + m\Omega^*| \leq \Delta$.

$$\mathcal{L}_{\text{rot}} = 1 - \frac{C_{\text{ring}} + C_{\text{rot}}}{2}.$$

Under mild narrow-band and window assumptions, $\mathcal{L}_{\text{rot}}$ serves as a slice-consistent differentiable surrogate for the

rotational SIM(2) constraint, and is bounded by the corresponding unified SIM(2) rotational-slice residual up to window/interpolation terms (proof in App. A.5). Compared with prior ‘‘temporal-peak’’ heuristics, our line-ratio with energy-weighted Ω^* suppresses non-rotational periodicities while remaining differentiable. Implementation details (polar resampling, $m=0$ exclusion, Δ , and weighting) are in App. D.2 and App. A.6.

3.4. Scaling Motion Loss

Scaling leaves a distinctive spectral signature: *radial energy flow and tilted lines* in the (ν, ω_t) plane obeying $\omega_t + \alpha \nu = 0$, i.e., the scaling slice of the unified SIM(2) hyperplane (cf. Eq. (1)). To keep the main text lightweight yet theory-linked, we adopt two robust proxies and state their consistency with the SIM(2) slice; the closed-form α^* and a tilted-line ratio C_{scale} are given in App. A.4–A.5.

Let $E_k(t)$ be the ring energy on the k -th annulus (App. D.2). We define $E_k(t)$ and the corresponding unit fields:

$$\nabla_\rho E_{k,t} = E_{k+1,t} - E_{k,t}, \quad \nabla_t E_{k,t} = E_{k,t+1} - E_{k,t},$$

$$\hat{\nabla}_\rho E = \frac{\nabla_\rho E}{\sqrt{\sum_{k,t} |\nabla_\rho E_{k,t}|^2 + \varepsilon}},$$

$$\hat{\nabla}_t E = \frac{\nabla_t E}{\sqrt{\sum_{k,t} |\nabla_t E_{k,t}|^2 + \varepsilon}}.$$

The (direction-agnostic) alignment score is

$$C_{\text{flow}} = \left| \sum_{k,t} \hat{\nabla}_\rho E_{k,t} \cdot \hat{\nabla}_t E_{k,t} \right|, \quad \rho_c(t) = \frac{\sum_k k E_k(t)}{\sum_k E_k(t) + \varepsilon_{\text{stab}}},$$

Let $\rho_c(t)$ denote the radial spectral centroid and define a bounded trend strength via correlation:

$$S_{\text{trend}} = |\text{corr}(\rho_c, t)| = \frac{|\text{cov}(\rho_c, t)|}{\sqrt{\text{var}(\rho_c) \text{var}(t) + \varepsilon}}.$$

$(E_k(t))$ is the spectral energy on the k -th concentric ring in the spatial-frequency plane (optionally normalized per t); $\varepsilon, \varepsilon_{\text{stab}} > 0$ are small positive numerical-stability constants.)

$$\mathcal{L}_{\text{scale}} = 1 - \frac{C_{\text{flow}} + S_{\text{trend}}}{2}.$$

Under a log-polar narrow-band shift model $E(i, t) \approx A(i - u(t))$ with $u'(t) = \alpha$ and mild window/interp errors, $\mathcal{L}_{\text{scale}}$ serves as a slice-consistent differentiable surrogate for the scaling SIM(2) constraint, and is bounded by the corresponding unified SIM(2) scaling-slice residual up to narrow-band, window, and interpolation terms (analysis and the closed-form α^* with the (ν, ω_t) tilted-line ratio C_{scale} in App. A.4–A.5). For very short windows ($T < 3$) we default both proxies to 0.5.

Table 1. Evaluation on the Physics Generation Benchmark (Meng et al., 2025). We compare Open-Sora with the same backbone trained using our frequency-domain motion regularizer; higher scores indicate better physical plausibility.

Metric	Mechanics $\uparrow$	Optics $\uparrow$	Thermal $\uparrow$	Material $\uparrow$	Average $\uparrow$
Open-Sora	0.37	0.44	0.46	0.50	0.44
Ours	0.45	0.55	0.54	0.55	0.52

3.5. Adaptive Weighting and Composite Loss

To accommodate different videos that may contain multiple motion patterns, we introduce an adaptive weighting mechanism based on a temperature parameter τ : $w_i^{\text{type}} = \frac{\exp(-\mathcal{L}_i/\tau)}{\sum_{j \in \mathcal{M}} \exp(-\mathcal{L}_j/\tau)}$. The theoretical foundation of this mechanism is grounded in the maximum-entropy principle from information theory. This gives higher weights to motion types with lower loss values (i.e., better conforming to specific motion patterns). When τ approaches 0, the weight distribution becomes more "winner-takes-all," highlighting the optimal motion type; higher τ values produce a smoother weight distribution, suitable for mixed motion. The final composite loss is the weighted sum of the motion losses:

$$\mathcal{L}_{\text{motion}} = \sum_{i \in \mathcal{M}} w_i^{\text{type}} \cdot \mathcal{L}_i. \quad (4)$$

This design not only enables the model to automatically identify the dominant motion type but also maintains sensitivity to mixed motion. Compared to traditional hard classification or single motion assumption methods, this approach better handles complex motion scenarios in the real world.

3.6. SIM(2)-Approximability of OpenVID-1M Clips

Since our loss targets projected SIM(2)-like apparent motions, we estimate how frequently such motions occur in OpenVID-1M. We randomly sample 1,000 training clips, track feature points across consecutive frames, and fit a global SIM(2) transformation with RANSAC (Fischler & Bolles, 1981). A clip is classified as SIM(2)-approximable if at least 60% of its frame pairs achieve an inlier ratio of at least 0.60 and a median reprojection error of at most 2.0 pixels.

Under these thresholds, 605 out of 1,000 clips, or 60.5%, are SIM(2)-approximable, with a mean inlier ratio of 0.686 and a median reprojection error of 0.57 pixels. This suggests that projected SIM(2)-like apparent motion is present in a substantial fraction of the training distribution. For clips that are not well described by SIM(2), the observability gates and adaptive weighting reduce the influence of unmatched motion branches.

Table 2. Quantitative evaluation on OpenVID-1M for two video diffusion backbones. We compare each baseline with the same model trained using our motion-aware regularizer; best results are in **bold**.

Metric	Open-Sora		MVDIT	
	Baseline	Ours	Baseline	Ours
VQA.A $\uparrow$	65.15	69.20	66.65	69.30
VQA.T $\uparrow$	59.57	69.71	63.96	70.24
SD Score $\uparrow$	68.24	68.42	68.31	68.78
CLIP Temporal Score $\uparrow$	99.80	99.85	99.83	99.89
Warping Error $\downarrow$	0.0089	0.0056	0.0080	0.0062
Temporal Consistency $\uparrow$	61.45	63.82	62.21	64.73
Action Recognition $\uparrow$	60.77	69.71	62.34	69.70
Motion Accuracy $\uparrow$	44.00	49.00	44.00	51.00
Flow Score $\uparrow$	1.15	1.18	1.01	1.22
Text-Video Alignment $\uparrow$	54.02	61.05	61.04	63.62
BLIP-BLEU $\uparrow$	23.73	24.52	24.14	24.76

4. Experimental Results

We evaluate whether the proposed frequency-domain regularizer improves physical motion while preserving visual quality and text-video alignment. Experiments are conducted on multiple video diffusion backbones, including Open-Sora, MVDIT, HunyuanVideo, and Wan 2.1-14B. We report standard video-generation metrics, physics-oriented benchmark scores, component ablations, human preference results, and qualitative comparisons.

4.1. Experimental Setup

Datasets and Models. We conducted our experiments using the OpenVID-1M dataset, a large-scale open-domain video dataset containing diverse motion patterns. For baseline models, we selected four video diffusion models: Open-Sora (Zheng et al., 2024), MVDIT (Nan et al., 2025), Hunyuan (Kong et al., 2024) and Wan 2.1 14B (Wan Team, 2025) which represent different architectural approaches to video diffusion.

Implementation details. We fine-tune each backbone for four epochs using a cosine-annealed LR initialized at 2×10^{-5} on 4 NVIDIA A100 GPUs. At every diffusion step t , we reconstruct the predicted clean sample $\hat{x}_0$, evaluate the physics-informed frequency loss on $\hat{x}_0$, and add it to the standard denoising objective. For an ϵ -prediction model, we use

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t, c)}{\sqrt{\bar{\alpha}_t}},$$

with the analogous clean-sample reconstruction for other parameterizations. Since $\hat{x}_0$ is less reliable at high noise levels, we use a linear decay weighting $w = 1 - t/T_{\text{diff}}$ to down-weight the physics loss at high noise levels. The motion loss is computed on a low-frequency 3D FFT cube with per-dimension cutoff $\varrho = 0.3$, which empirically retains 97.5% of spectral energy on 1k random videos; see App. C. We use

a temporal FFT window of $K = 16$ frames. The weighted least-squares block is fully differentiable, runs in FP32 with autocast off, and uses ridge $\lambda=10^{-3}$ and $\varepsilon=10^{-8}$ (rest in BF16); see App. A.6–C for windows, polar/log resampling, gating and stability details.

Evaluation protocol. Following EvalCrafter (Liu et al., 2024b) and the OpenVID-1M setting, we report four categories of metrics to enable fair comparison on this benchmark: *visual quality* (VQA_A, VQA_T, SD-Score), *temporal coherence* (CLIP-Temporal, Warping Error, Temporal Consistency), *motion quality* (Action Recognition, Motion Accuracy, Flow), and *text alignment* (Text–Video Alignment, BLIP-BLEU). We use the official EvalCrafter implementation and evaluation protocol for OpenVID-1M to ensure comparability with prior work.

4.2. Quantitative Results

Comparison on Open-Sora. Table 2 presents our quantitative evaluation results on the Open-Sora model. Our approach consistently outperforms the baseline across most metrics. Notably, we observe substantial improvements in motion-related metrics, with a 8.9%-point improvement in Action Recognition Score and a 5%-point improvement in Motion Accuracy Score. This confirms that our frequency domain-based approach effectively enhances the physical plausibility of motion patterns in generated videos.

We further evaluate on a recent Physics Generation Benchmark (Meng et al., 2025) that isolates mechanics, optics, thermal, and material phenomena. Our method improves the average physics score from 0.44 to 0.52 over Open-Sora (Table 1), suggesting improved performance on physics-oriented evaluation metrics.

Comparison on MVDIT. Table 2 shows similar improvements on the MVDIT model, indicating that our approach generalizes well across different video diffusion architectures. The consistent performance gains across both models validate the effectiveness of our physics-informed frequency domain approach.

Importantly, our method achieves these motion quality improvements without sacrificing visual quality or semantic alignment. We observe modest improvements in these dimensions as well, suggesting that enhancing physical motion plausibility can positively influence overall video generation quality.

Comparison on Hunyuan (Kong et al., 2024) (LoRA). For the *Hunyuan* model, we adopt LoRA (PEFT) due to its size, inserting low-rank adapters into attention and time-related linear layers while freezing all other weights. To ensure fairness, the baseline and “+Ours” use identical LoRA placement, rank, and training schedule. Under this fixed adapter budget, our frequency-domain physical loss yields

Table 3. Evaluation on HunyuanVideo (Kong et al., 2024) under matched LoRA fine-tuning. We compare the baseline, base LoRA fine-tuning, a flow-based temporal-consistency reference, and our motion-regularized LoRA variant.

Metric	Hunyuan Baseline	LoRA Base loss	Flow Loss	LoRA Ours Loss
VQA_A $\uparrow$	73.85	73.84	73.92	74.79
VQA_T $\uparrow$	85.14	85.16	85.36	87.37
SD Score $\uparrow$	68.32	68.42	68.37	69.83
CLIP Temporal Score $\uparrow$	99.91	99.90	99.92	99.95
Warping Error $\downarrow$	0.0024	0.0022	0.0022	0.0016
Temporal Consistency $\uparrow$	63.65	63.70	64.23	66.03
Action Recognition Score $\uparrow$	68.93	69.02	68.78	73.15
Motion Accuracy Score $\uparrow$	56.00	56.0	56.0	59.0
Flow Score $\uparrow$	1.39	1.39	1.45	1.46
Text-Video Alignment $\uparrow$	59.60	60.62	60.33	65.34
BLIP-BLEU $\uparrow$	24.41	24.42	24.60	25.23

Table 4. Stratified evaluation by motion complexity. Test prompts are split into simple and complex motion subsets.

Metric	Baseline complex	Ours complex	Baseline simple	Ours simple
SD Score $\uparrow$	70.20	71.34	67.82	69.42
CLIP Temporal Score $\uparrow$	99.92	99.96	99.90	99.95
Warping Error $\downarrow$	0.0020	0.0011	0.0025	0.0018
Flow Score $\uparrow$	1.39	1.46	1.38	1.46
BLIP-BLEU $\uparrow$	26.41	26.92	23.87	24.77

consistent gains in *Motion Quality* and maintains (or slightly improves) *Visual Quality* and *Text Alignment* (see Table 3).

Additional evaluation on Wan 2.1-14B (Wan Team, 2025) (LoRA). To test whether the proposed method remains effective on a stronger recent video generator, we further evaluate it on Wan 2.1-14B with LoRA finetuning. As shown in Table 6, our method improves multiple metrics over the no-motion-loss baseline, including a +8.47 gain in Visual Quality.

We also evaluate Wan 2.1-14B on PhyGenBench to examine whether the improvements extend beyond generic video-quality metrics. As shown in Table 5, the full model improves all four categories. The component ablations suggest distinct category-wise dependencies: Mechanics is most sensitive to the translation component, Optics to the rotation component, and Thermal to the scaling component, while Material shows more moderate sensitivity across components. We further evaluate sensitivity to the weighting temperature τ and low-pass ratio ϱ on Wan 2.1-14B. Across all tested settings, the motion loss improves visual quality over the no-motion-loss baseline, with the default setting achieving the best overall balance; see App. D.4.

Comparison with MotionCraft. For completeness, we include a reference comparison with MotionCraft (Savant Aira et al., 2024), a zero-shot method that uses externally simulated optical flow to warp the latent space of a frozen image diffusion model. Our method instead regularizes motion



Figure 3. Comparison between ours and the baseline (Hunyuan) under translation, rotation, and scaling.

Table 5. PhyGenBench results and component ablations on Wan 2.1-14B. We report the full motion loss and variants with the translation, rotation, or scaling component removed; higher scores indicate better physical plausibility.

Setting	Mechanics $\uparrow$	Optics $\uparrow$	Thermal $\uparrow$	Material $\uparrow$	Average $\uparrow$
No motion loss	0.517	0.707	0.433	0.533	0.548
Full (ours)	0.600	0.747	0.500	0.560	0.602
w/o translation	0.533	0.747	0.444	0.520	0.561
w/o rotation	0.567	0.693	0.478	0.547	0.571
w/o scaling	0.592	0.720	0.422	0.547	0.570

Table 6. Evaluation on Wan 2.1-14B with LoRA finetuning.

Method	VQA-A $\uparrow$	VQA-T $\uparrow$	IS $\uparrow$	SD-Score $\uparrow$	Vis. Qual. $\uparrow$
No motion loss	67.48	64.43	16.28	67.82	54.18
Ours	71.18	66.98	18.09	68.43	62.65

during training in the frequency domain. Using the officially released MotionCraft examples, our method achieves better results on all four metrics in Table 7.

Flow-based Temporal Consistency (Strong Baseline).

While optical flow has long been used to impose temporal coherence via warping-based photometric losses in video generation and processing, most contemporary text-to-video diffusion systems do not include an explicit flow-matching loss during training; they typically rely on architectural inductive biases, cross-frame conditioning, or inference-time alignment. For completeness, we implement a standard flow-based consistency loss. Given adjacent frames x_t, x_{t+1} , we estimate forward/backward optical flow $\mathbf{F}_{t \rightarrow t+1}, \mathbf{F}_{t+1 \rightarrow t}$ with RAFT (Teed & Deng, 2020). Let $\mathcal{W}$ be bilinear warping and $\mathbf{M}, \mathbf{M}'$ be visibility masks. Our loss is a masked reprojection error with robust Charbonnier/ ℓ_1 penalty plus a standard flow smoothness term:

$$\mathcal{L}_{\text{flow}} = \|\mathbf{M} \odot (x_{t+1} - \mathcal{W}(x_t, \mathbf{F}_{t \rightarrow t+1}))\|_1 + \|\mathbf{M}' \odot (x_t - \mathcal{W}(x_{t+1}, \mathbf{F}_{t+1 \rightarrow t}))\|_1 + \lambda \mathcal{L}_{\text{smooth}}(\mathbf{F}).$$

This provides a temporal-consistency regularizer that di-

rectly constrains cross-frame correspondences, is sensitive to jitter/stretching artifacts, needs no backbone changes, and adds controllable overhead (RAFT inference only). As shown in Table 3, Our model outperforms this baseline.

User Study and Analysis. We adopted the Two-Alternative Forced-Choice (2AFC) protocol to evaluate whether our design improves different backbone models. For each trial, participants were shown two videos side-by-side—one from a vanilla baseline model, the other from the same model augmented with ours and asked to choose which looked better overall. We integrated our approach into representative state-of-the-art video generation models and pooled all comparisons together. A total of 106 participants each completed 15 randomly ordered trials, with left-right placement of baseline vs. augmented outputs independently randomized to eliminate positional bias. Viewers were instructed to consider both visual quality (sharpness, color fidelity) and motion naturalness (motion smoothness, coherence) in the first question.

As shown in Figure 6, our augmented models were preferred over their vanilla counterparts across every backbone, with overall preference rates ranging from 74.2 % to 82.7 % depending on the architecture. The user study interface is shown in Figure 4. These results demonstrate that our method not only boosts performance on a single model, but also generalizes effectively across multiple distinct video generation pipelines.

Stratified evaluation by motion complexity. To isolate the effect of our physics-guided motion loss, we stratify test prompts into simple vs. complex motion using an LLM (GPT-5) following our rubric: a prompt is simple when the dominant motion is well-approximated by a single rigid transform (translation, rotation, scaling); otherwise it is complex. The complete split is provided in the Supplementary. We then recompute metrics for the baseline and baseline + ours within each stratum. As shown in Table 4, we stratify by motion complexity and observe consistent improvements

Table 7. Reference comparison with MotionCraft. Although the settings differ, our method performs better on all four metrics.

Method	Frame Consistency $\uparrow$	Motion Consistency $\uparrow$	LPIPS flow $\downarrow$	Warping Error $\downarrow$
Ours	0.9988	0.9468	0.0824	0.0045
MotionCraft	0.9943	0.8563	0.1250	0.0104

over the baseline in both the simple and complex subsets. On average, gains are larger on the simple subset.

4.3. Ablation Studies

Component ablation. In our ablation study on the OpenSora (Zheng et al., 2024) model (Table 8), we find that each motion-specific loss term contributes uniquely to overall performance. Omitting the translation loss leads to noticeable drops in both visual quality and basic motion fidelity, while removing the rotation loss mainly affects semantic alignment and cyclical motion consistency. Excluding the scaling loss produces the smallest overall degradation but still measurably worsens zoom-related coherence and motion accuracy. Importantly, none of the individual removals recovers the balanced improvements achieved by the full composite loss, suggesting that translation, rotation, and scaling constraints act synergistically to yield the best trade-off across visual, temporal, and motion-quality metrics.

Timestep weighting and temporal window. We use a linear timestep decay $w = 1 - t/T_{\text{diff}}$ to down-weight the physics loss at high noise levels, where the predicted clean sample $\hat{x}_0$ is less reliable. We use the same rule across all evaluated backbones without backbone-specific tuning. Ablations show that applying the physics loss uniformly across all timesteps substantially degrades performance, while the linear decay gives the best overall balance. We also use a fixed temporal FFT window of $K = 16$ frames as a compromise between slow- and fast-motion sensitivity; shorter windows favor fast motion, whereas longer windows better capture slow motion. Additional ablations of timestep weighting and temporal-window length are provided in App. D.5.

We additionally test the robustness of our spectral diagnostics to brightness changes and common rendering perturbations. The matched loss changes only slightly under brightness, blur, noise, and JPEG perturbations, and the selected motion type remains correct in every case. These results suggest that the frequency-domain diagnostics are not overly sensitive to simple appearance-level perturbations; details are provided in App. D.3.

4.4. Qualitative Analysis

Fig. 3 contrasts our method with the Hunyuan baseline on three canonical motions. (i) Translation. For a football

Table 8. Ablation study of individual motion loss components on the OpenVID-1M dataset using the Open-Sora model. We report the full loss and variants where the translation loss, rotation loss, or scaling motion loss is removed.

Metric	Full	w/o Trans.	w/o Rot.	w/o Scale
VQA-A $\uparrow$	69.20	66.15	66.95	67.10
VQA-T $\uparrow$	69.71	64.80	65.00	65.40
SD Score $\uparrow$	68.42	67.85	68.10	68.20
CLIP Temporal Score $\uparrow$	99.85	99.70	99.75	99.80
Warping Error $\downarrow$	0.0056	0.0078	0.0070	0.0068
Temporal Consistency $\uparrow$	63.82	61.20	62.10	62.18
Action Recognition Score $\uparrow$	69.71	66.20	66.50	67.00
Motion Accuracy Score $\uparrow$	49.00	44.00	45.00	44.00
Flow Score $\uparrow$	1.18	1.15	1.16	1.16
Text-Video Alignment $\uparrow$	61.05	57.60	58.00	58.20
BLIP-BLEU $\uparrow$	24.52	23.80	23.92	23.43

prompted to move right-to-left, the baseline remains nearly static for $\sim 80\%$ of the sequence, whereas our method produces smooth, monotonic displacement. (ii) Clockwise rotation. The baseline reverses direction mid-sequence (left then right), breaking temporal consistency; our method maintains a single, persistent clockwise rotation. (iii) Forward motion with scale. For a train advancing toward the camera, the baseline shows abrupt appearance of the train with jerky forward motion, while ours yields smooth central approach with the expected increase in scale (see supplementary video). These qualitative observations are consistent with our quantitative results and indicate that our frequency-domain formulation better preserves directionality, phase, and scale dynamics across motion types. These examples correspond to baseline failures under simple projected SIM(2) prompts that our method corrects; additional examples are included in the supplementary.

5. Discussion and Conclusion

Limitations and future work. Our framework currently addresses SIM(2) and projected SIM(2)-like apparent motions in the image plane, including translation, in-plane rotation, and uniform scaling. These patterns can arise from simple object motion, camera-induced motion such as pan/tilt, roll, and zoom/dolly, or scene elements whose projections are locally approximated by these transformations. However, the current formulation does not explicitly model more complex physical dynamics such as elastic deformation, articulated movement, fluids, or contact-rich interactions. Extending beyond these fundamental projected motion patterns is left for future work.

Conclusion. We have introduced a unified, frequency-domain approach for enforcing physical motion plausibility in video diffusion models. Our method consistently improves action recognition, optical flow quality, and motion accuracy while preserving visual fidelity and semantic alignment.

Impact Statement

This work aims to advance machine learning methods for improving the physical plausibility and temporal consistency of video generation models. Potential positive impacts include more reliable synthetic video generation for creative tools, simulation, education, and visual-content creation. We do not identify any additional societal risks beyond those generally associated with video generation models.

References

- Adelson, E. H. and Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, Feb 1985. doi: 10.1364/JOSAA.2.000284. URL <https://opg.optica.org/josaa/abstract.cfm?URI=josaa-2-2-284>.
- 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, 2023. doi: 10.1109/CVPR52729.2023.02161.
- Bracewell, R. N. Strip integration in radio astronomy. *Australian Journal of Physics*, 9:198–217, 1956. URL <https://api.semanticscholar.org/CorpusID:121023618>.
- Brigham, E. O. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., USA, 1988. ISBN 0133075052.
- Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- Cai, Y., Li, K., Jia, M., Wang, J., Sun, J., Liang, F., Chen, W., Juefei-Xu, F., Wang, C., Thabet, A., Dai, X., Ju, X., Yuille, A., and Hou, J. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation, 2026. URL <https://arxiv.org/abs/2512.24551>.
- Dong, D. W. and Atick, J. J. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6(3):345, aug 1995. doi: 10.1088/0954-898X/6/3/003. URL <https://doi.org/10.1088/0954-898X/6/3/003>.
- Fischler, M. A. and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL <https://doi.org/10.1145/358669.358692>.
- Fleet, D. J. and Jepson, A. D. Computation of component image velocity from local phase information. *Int. J. Comput. Vision*, 5(1):77–104, September 1990. ISSN 0920-5691. doi: 10.1007/BF00056772. URL <https://doi.org/10.1007/BF00056772>.
- Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.
- Hartley, R. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.
- Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., and Wood, F. Flexible diffusion modeling of long videos. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- Kataoka, H., Wakamiya, T., Hara, K., and Satoh, Y. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*, 2020.
- Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Kumar, M., Babaeizadeh, M., Erhan, D., Finn, C., Levine, S., Dinh, L., and Kingma, D. Videoflow: A conditional flow-based model for stochastic video generation. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rJgUfTEYvH>.
- Le Guen, V. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11471–11481, 2020. doi: 10.1109/CVPR42600.2020.01149.
- Lindeberg, T. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, USA, 1993. ISBN 0792326369.

- Liu, S., Ren, Z., Gupta, S., and Wang, S. Physgen: Rigid-body physics-grounded image-to-video generation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXII*, pp. 360–378, Berlin, Heidelberg, 2024a. Springer-Verlag. ISBN 978-3-031-73006-1. doi: 10.1007/978-3-031-73007-8_21. URL https://doi.org/10.1007/978-3-031-73007-8_21.
- Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., and Shan, Y. Evalcrafter: Benchmarking and evaluating large video generation models, 2024b. URL <https://arxiv.org/abs/2310.11440>.
- Ma, G., Huang, H., Yan, K., Chen, L., Duan, N., Yin, S., Wan, C., Ming, R., Song, X., Chen, X., Zhou, Y., Sun, D., Zhou, D., Zhou, J., Tan, K., An, K., Chen, M., Ji, W., Wu, Q., Sun, W., Han, X., Wei, Y., Ge, Z., Li, A., Wang, B., Huang, B., Wang, B., Li, B., Miao, C., Xu, C., Wu, C., Yu, C., Shi, D., Hu, D., Liu, E., Yu, G., Yang, G., Huang, G., Yan, G., Feng, H., Nie, H., Jia, H., Hu, H., Chen, H., Yan, H., Wang, H., Guo, H., Xiong, H., Xiong, H., Gong, J., Wu, J., Wu, J., Wu, J., Yang, J., Liu, J., Li, J., Zhang, J., Guo, J., Lin, J., Li, K., Liu, L., Xia, L., Zhao, L., Tan, L., Huang, L., Shi, L., Li, M., Li, M., Cheng, M., Wang, N., Chen, Q., He, Q., Liang, Q., Sun, Q., Sun, R., Wang, R., Pang, S., Yang, S., Liu, S., Liu, S., Gao, S., Cao, T., Wang, T., Ming, W., He, W., Zhao, X., Zhang, X., Zeng, X., Liu, X., Yang, X., Dai, Y., Yu, Y., Li, Y., Deng, Y., Wang, Y., Wang, Y., Lu, Y., Chen, Y., Luo, Y., Luo, Y., Yin, Y., Feng, Y., Yang, Y., Tang, Z., Zhang, Z., Yang, Z., Jiao, B., Chen, J., Li, J., Zhou, S., Zhang, X., Zhang, X., Zhu, Y., Shum, H.-Y., and Jiang, D. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. URL <https://arxiv.org/abs/2502.10248>.
- 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 *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=dIjMswSzgF>.
- Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., and Tai, Y. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=j7kdXSrISM>.
- Oppenheim, A. V., Schaffer, R. W., and Buck, J. R. *Discrete-time signal processing (2nd ed.)*. Prentice-Hall, Inc., USA, 1999. ISBN 0137549202.
- Reddy, B. and Chatterji, B. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996. doi: 10.1109/83.506761.
- Ruderman, D. L. and Bialek, W. Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.*, 73:814–817, Aug 1994. doi: 10.1103/PhysRevLett.73.814. URL <https://link.aps.org/doi/10.1103/PhysRevLett.73.814>.
- Saito, M., Matsumoto, E., and Saito, S. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.
- Savant Aira, L., Montanaro, A., Aiello, E., Valsesia, D., and Magli, E. Motioncraft: Physics-based zero-shot video generation. In *Advances in Neural Information Processing Systems*, 2024.
- Sevilla-Lara, L., Sun, D., Jampani, V., and Black, M. J. Optical flow with semantic segmentation and localized layers. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3889–3898, 2016. doi: 10.1109/CVPR.2016.422.
- Sharma, U. and Duits, R. Left-invariant evolutions of wavelet transforms on the similitude group. *Applied and Computational Harmonic Analysis*, 39(1):110–137, 2015. ISSN 1063-5203. doi: <https://doi.org/10.1016/j.acha.2014.09.001>. URL <https://www.sciencedirect.com/science/article/pii/S1063520314001171>.
- Simoncelli, E. P. and Heeger, D. J. A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–761, 1998. ISSN 0042-6989. doi: [https://doi.org/10.1016/S0042-6989\(97\)00183-1](https://doi.org/10.1016/S0042-6989(97)00183-1). URL <https://www.sciencedirect.com/science/article/pii/S0042698997001831>.
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., and Taigman, Y. Make-a-video: Text-to-video generation without text-video data, 2022. URL <https://arxiv.org/abs/2209.14792>.
- Stephane, M. A wavelet tour of signal processing, 1999.
- Teed, Z. and Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, pp. 402–419, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58535-8. doi: 10.1007/978-3-030-58536-5_24. URL https://doi.org/10.1007/978-3-030-58536-5_24.

- Tesfaldet, M., Brubaker, M. A., and Derpanis, K. G. Two-stream convolutional networks for dynamic texture synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. Mocogan: Decomposing motion and content for video generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1526–1535, 2018. doi: 10.1109/CVPR.2018.00165.
- Villegas, R., Pathak, A., Kannan, H., Erhan, D., Le, Q. V., and Lee, H. *High fidelity video prediction with large stochastic recurrent neural networks*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- Vondrick, C., Pirsiaavash, H., and Torralba, A. Generating videos with scene dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, pp. 613–621, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.
- Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- Wang, J. Y. and Adelson, E. H. Representing moving images with layers. *Trans. Img. Proc.*, 3(5):625–638, September 1994. ISSN 1057-7149. doi: 10.1109/83.334981. URL <https://doi.org/10.1109/83.334981>.
- Watson, A. B. and Ahumada, A. J. Model of human visual-motion sensing. *J. Opt. Soc. Am. A*, 2(2):322–342, Feb 1985. doi: 10.1364/JOSAA.2.000322. URL <https://opg.optica.org/josaa/abstract.cfm?URI=josaa-2-2-322>.
- Weissenborn, D., Täckström, O., and Uszkoreit, J. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rJgsskrFwH>.
- Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. T. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, August 2019. ISSN 0920-5691. doi: 10.1007/s11263-018-01144-2. URL <https://doi.org/10.1007/s11263-018-01144-2>.
- Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., Fu, J., Gong, M., Wang, L., Liu, Z., Li, H., and Duan, N. NUWA-XL: Diffusion over diffusion for eXtremely long video generation. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1309–1320, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.73. URL <https://aclanthology.org/2023.acl-long.73/>.
- Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

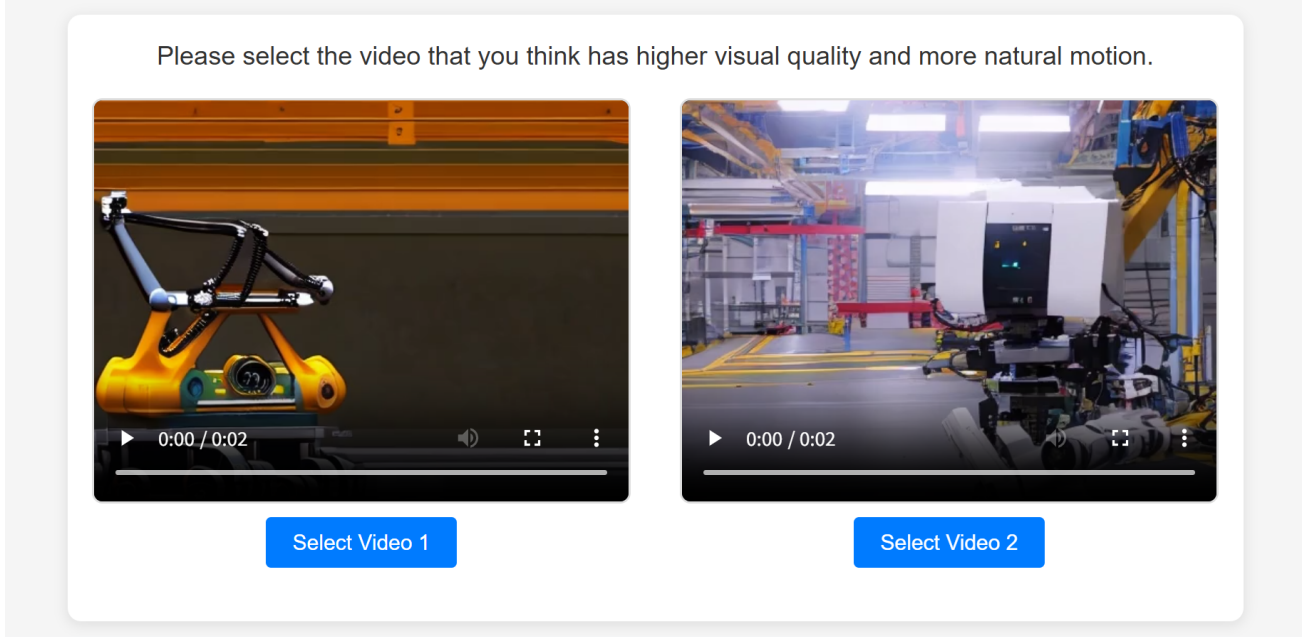


Figure 4. User study interface design. Our evaluation employs a two-alternative forced choice (2AFC) protocol with two types of questions. The first type (shown) asks participants to select the video with higher visual quality and more natural motion from a pair of generated videos displayed side-by-side. The second type evaluates text-video alignment by asking participants to choose which video better matches the given text prompt. Video positions are randomized to eliminate positional bias.

A. Appendix

A.1. Notation and Transforms

We denote the input video window by $V \in \mathbb{R}^{T \times H \times W}$ (one channel for simplicity; RGB is handled channel-wise with energies summed). We work in the complex spatiotemporal Fourier domain. Let

$$\hat{V}(\omega_x, \omega_y, \omega_t) = \text{DFT}_t \left\{ h[t] \cdot \text{DFT}_{x,y} \left\{ V(t, \cdot, \cdot) - \frac{1}{2} \right\} \right\},$$

be the separable 2D spatial DFT per frame followed by a 1D temporal DFT (with an optional Hann window $h[t]$); frequencies are indexed by $(\omega_x, \omega_y, \omega_t)$. On the spatial frequency plane we adopt polar coordinates (ρ, θ) with $\rho = \sqrt{\omega_x^2 + \omega_y^2}$ and $\theta = \text{atan2}(\omega_y, \omega_x)$. Angular harmonics are obtained via a DFT over θ , yielding $C_m(\rho, \omega_t)$ indexed by $m \in \mathbb{Z}$; log-radial harmonics are obtained by re-sampling ρ on a logarithmic grid $\xi = \log \rho$ and applying a 1D DFT to obtain $D_\nu(\omega_t)$ indexed by $\nu \in \mathbb{Z}$. We compute energies as $E(\cdot) = |\hat{V}(\cdot)|^2$.

A.2. Unified SIM(2) Spectral Manifold: Derivation

We consider a short temporal window where the dominant motion is well-approximated by a similarity transform: translation $v = (v_x, v_y)$, in-plane rotation with angular velocity Ω , and isotropic scaling rate $\alpha = \dot{\sigma}$ (with $s(t) = e^{\sigma(t)}$).

Translation. For $V(x, y, t) = V_0(x - v_x t, y - v_y t)$, the DFT analysis gives the classical plane constraint

$$\omega_t + v_x \omega_x + v_y \omega_y = 0. \quad (5)$$

Rotation. Write $V(r, \theta, t) = V_0(r, \theta - \Omega t)$. Expanding V_0 in angular harmonics and transforming in t yields

$$\hat{V}(\rho, \theta, \omega_t) = \sum_{m \in \mathbb{Z}} e^{im\theta} \mathcal{B}_m(\rho) \delta(\omega_t + m\Omega), \quad (6)$$

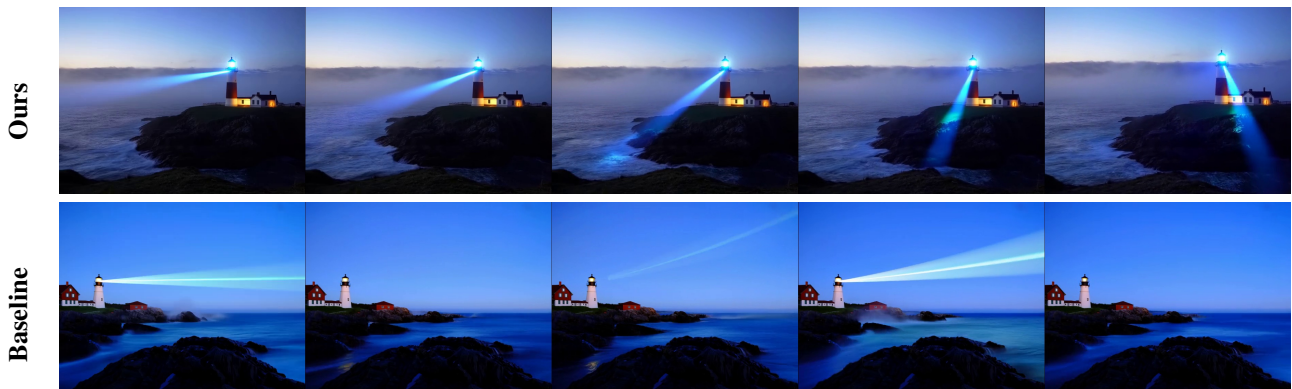
where $\mathcal{B}_m$ involves Bessel kernels (annular concentration). Hence energy concentrates along *tilted lines* $\omega_t + \Omega m = 0$ in the (m, ω_t) plane. **Scaling.** Let $V(x, y, t) = V_0(\frac{x}{s(t)}, \frac{y}{s(t)})$ with $s(t) = e^{\sigma(t)}$. In spatial frequency, scaling is a dilation; in



Prompt: A rocket performs a controlled vertical landing onto a coastal pad.



Prompt: A freight train arcs through a canyon.



Prompt: A lighthouse's rotating beacon sweeps; projected light cones translate and scale over rocks and waves.

Figure 5. Additional comparisons with baseline (Hunyuan). In the first example, the baseline's rocket lands and then takes off again. In the second example, the train car colors change abruptly, and additional train cars appear out of nowhere. In the third example, the baseline's lighthouse light is inconsistent and doesn't generate according to the prompt instructions. However, our motion is all coherent.

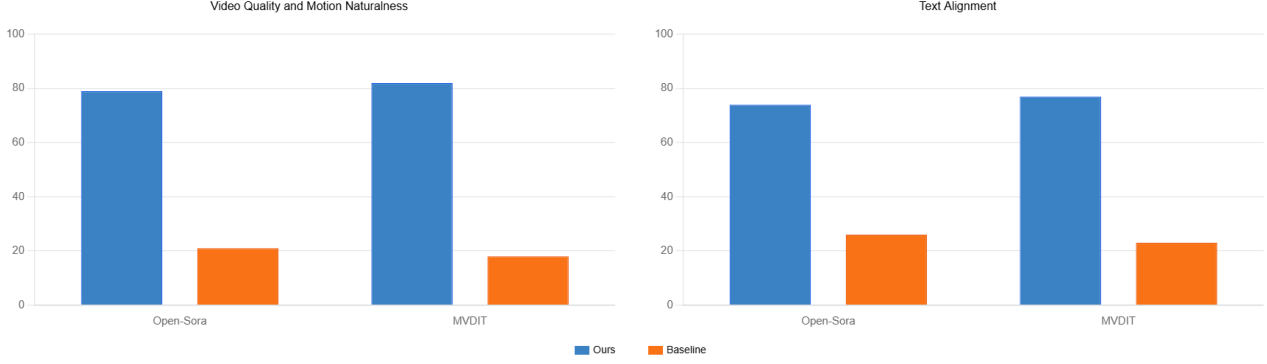


Figure 6. User study results. We compare our method to two baselines (Open-Sora and MVDIT) using a two-alternative forced choice protocol. For each baseline, we report the percentage of user votes in our favor (blue) and in favor of the baseline (orange). Our method was consistently preferred by users in both video quality/motion naturalness evaluation and text-prompt alignment assessment, achieving preference rates of 79.4% and 74.2% against Open-Sora, and 82.7% and 77.9% against MVDIT, respectively.

log-radius $\xi = \log \rho$, it becomes a shift $\xi \mapsto \xi - \sigma(t)$. Therefore the (ν, ω_t) spectrum concentrates on

$$\omega_t + \alpha \nu = 0, \quad \alpha = \dot{\sigma}. \quad (7)$$

Unified hyperplane. Collecting the three facts above, an ideal SIM(2) motion concentrates spectral energy on a single hyperplane in $(\omega_x, \omega_y, m, \nu, \omega_t)$:

$$\omega_t + v_x \omega_x + v_y \omega_y + \Omega m + \alpha \nu + b_0 = 0, \quad (8)$$

where b_0 absorbs constant phase/windowing terms. Eq. (5) and the rotational/scaling tilted lines are recovered by setting the other coefficients to zero.

A.3. Gold-Standard Residual and Basic Properties

From the observed spectrum we build weighted samples (ϕ_i, b_i, w_i) with

$$\phi_i = [\omega_x, \omega_y, m, \nu, 1], \quad b_i = -\omega_t, \quad w_i \geq 0,$$

and estimate $\theta = [v_x, v_y, \Omega, \alpha, b_0]^\top$ via weighted ridge regression:

$$\hat{\theta} = \arg \min_{\theta} \sum_i w_i (\phi_i \theta - b_i)^2 + \lambda \|\theta\|_2^2. \quad (9)$$

The unified *gold-standard residual* is

$$\mathcal{L}_{\text{uni}} = \frac{\sum_i w_i (\phi_i \hat{\theta} - b_i)^2}{\sum_i w_i}. \quad (10)$$

Proposition 1 (Zero-residual for ideal SIM(2)). If the video window follows an exact SIM(2) motion with constant parameters, then $\mathcal{L}_{\text{uni}} = 0$ up to boundary/windowing terms. *Sketch.* The spectral mass is supported on the hyperplane (8); any least-squares estimate lying on that plane gives zero orthogonal projection residual. Parseval’s identity transfers the spatial transformations to spectral constraints; see standard derivations for translation planes, rotational angular-harmonic $\delta(\omega_t + m\Omega)$ lines, and log-radius temporal shifts.

Proposition 2 (Consistency under noise). Assume i.i.d. additive spectral noise with finite second moment, weights bounded and bounded away from zero on a set of positive measure, and sufficient excitation in $(\omega_x, \omega_y, m, \nu)$. Then $\hat{\theta} \rightarrow \theta^*$ and $\mathcal{L}_{\text{uni}} \rightarrow \sigma^2$ (noise floor) as the number of samples increases. Weighted ridge ensures a bounded condition number.

Observability. The parameters are locally identifiable when the design matrix has full column rank in the energy-weighted sense, which requires non-degenerate support in $\{\omega_x, \omega_y\}$ for translation, nontrivial $|m|$ for rotation, and nontrivial $|\nu|$ for scaling.

A.4. From Theory to Computables: Rotation and Scaling Sufficient Statistics

Rotation. Define the angular DFT and its temporal DFT:

$$C_m(\rho, t) = \frac{1}{2\pi} \int_0^{2\pi} \widehat{V}(\rho, \theta, t) e^{-im\theta} d\theta, \quad \tilde{C}_m(\rho, \omega_t) = \text{DFT}_t\{C_m(\rho, t)\}.$$

An energy-weighted least-squares estimate for Ω has the closed-form

$$\Omega^* = -\frac{\sum_{\rho, m, \omega_t} |\tilde{C}_m|^2 \omega_t m}{\sum_{\rho, m, \omega_t} |\tilde{C}_m|^2 m^2}. \quad (11)$$

The *tilted-line energy ratio*

$$E_{\text{line}} = \sum_{\rho} \sum_{m \neq 0} \sum_{|\omega_t + m\Omega^*| \leq \Delta} |\tilde{C}_m|^2, \quad E_{\text{all}} = \sum_{\rho} \sum_{m \neq 0} \sum_{\omega_t} |\tilde{C}_m|^2, \quad C_{\text{rot}} = \frac{E_{\text{line}}}{E_{\text{all}}} \quad (12)$$

directly measures concentration along $\omega_t + \Omega m = 0$. Combined with the annular concentration C_{ring} , our rotation loss is $\mathcal{L}_{\text{rot}} = 1 - \frac{1}{2}(C_{\text{ring}} + C_{\text{rot}})$.

Scaling. After log-radius re-sampling $\xi = \log \rho$ with a 1D-DFT along ξ , we obtain $D_{\nu}(t)$ and its temporal DFT $\tilde{D}_{\nu}(\omega_t)$. The analogous estimate and tilted-line ratio are

$$\alpha^* = -\frac{\sum_{\nu, \omega_t} |\tilde{D}_{\nu}|^2 \omega_t \nu}{\sum_{\nu, \omega_t} |\tilde{D}_{\nu}|^2 \nu^2}, \quad C_{\text{scale}} = \frac{\sum_{|\omega_t + \alpha^* \nu| \leq \Delta} |\tilde{D}_{\nu}|^2}{\sum_{\omega_t, \nu} |\tilde{D}_{\nu}|^2}. \quad (13)$$

In the main paper we keep two simpler, robust proxies: radial-flow consistency C_{flow} and centroid trend S_{trend} . In §A.5 we relate these proxies to the unified residual on the scaling slice, showing that they are bounded by the corresponding unified residual up to narrow-band terms.

A.5. Relations to the Gold-Standard Residual

Let $\mathcal{L}_{\text{uni}}^{(\text{rot})}$ denote (10) restricted to the rotational slice (i.e., translation/scaling features clamped). Under an energy narrow-band assumption (dominant annulus/ring) and mild noise, there exist constants $a, b > 0$ such that

$$1 - \frac{1}{2}(C_{\text{ring}} + C_{\text{rot}}) \leq a \cdot \mathcal{L}_{\text{uni}}^{(\text{rot})} + b \cdot \varepsilon_{\text{NB}}. \quad (14)$$

A similar bound holds for scaling with C_{flow} and S_{trend} :

$$1 - \frac{1}{2}(C_{\text{flow}} + S_{\text{trend}}) \leq c_1 \cdot \mathcal{L}_{\text{uni}}^{(\text{scale})} + c_2 \cdot \varepsilon_{\text{NB}}. \quad (15)$$

Sketch. The tilted-line ratios are sufficient statistics for concentration in the (m, ω_t) or (ν, ω_t) plane; Cauchy–Schwarz yields that the off-line energy controls the orthogonal projection residual to the corresponding slice of the unified plane. The ring/flow terms compensate for spatial localization (annular narrow-bandness), with ε_{NB} capturing deviation from narrow-band energy.

For translation, the weighted plane-fitting residual used in the main paper is precisely a computable proxy for the projection residual onto Eq. (5), and hence matches the unified residual on the translation slice up to weighting and conditioning constants.

A.6. Implementation Details

Polar interpolation. We pre-compute a look-up table (LUT) to map polar bins (ρ_k, θ_{ℓ}) to Cartesian indices (ω_x, ω_y) and use bilinear interpolation on the spectral grid. We apply a Hann window in time before the temporal DFT to reduce leakage. Typical choices: $N_r \in \{16, 20\}$ rings (linear in ρ) and $M \in \{16, 24\}$ angular bins; for scaling, log-radius sampling uses $N_{\xi} \in \{16, 24\}$ bins.

Observability and energy gates. We use an energy gate $g_E = \sigma(f(\frac{E}{E_{\max}} - \tau_E))$ with $\tau_E \in [0.1, 0.2]$ and $f \in [6, 10]$, and an observability gate for rotation $g_{\text{obs}}(m) = \frac{m^2}{m^2 + \lambda}$ with $\lambda = 1$. The final weight is $w_i = g_E \cdot g_{\text{obs}}$ (and analogous for scaling with ν).

Robust regression and regularization. The unified regression is solved with ridge $\lambda = 10^{-3}$. In high-noise settings we replace squared loss by Huber/Charbonnier; the closed forms (11), (13) are retained for initialization, followed by one Gauss–Newton step.

Unmatched branches. For clips where a given SIM(2) slice is weakly supported, the energy/observability gates suppress low-excitation samples, and the robustified residuals avoid large uninformative gradients. Combined with the adaptive weighting, unsupported motion branches contribute little to the composite loss rather than forcing all clips to match all SIM(2) components.

A.7. Computational Cost

In our settings, wall-clock overhead is typically 10–20% of the backbone forward pass. Our method does not modify the backbone or sampler; inference compute is unchanged by design.

A.8. Reproducibility Checklist

- Hardware: NVIDIA A100 $\times 4$; mixed precision BF16.
- Training: cosine LR schedule, initial LR 2×10^{-5} , window $T = 12$ –16, loss weights as in the main paper.
- Spectral settings: rings $N_r=20$, angles $M=24$, log-radius bins $N_\xi=24$, tolerance $\Delta=1$ freq. bin, $\varepsilon=10^{-8}$.

B. Rigorous Relations Between Unified Spectral Residual and Surrogate Losses

B.1. Notation, Setting, and Assumptions

Spectral samples and design. For each discrete spectral sample i , define

$$\phi_i = [\omega_x, \omega_y, m, \nu, 1], \quad b_i = -\omega_t,$$

with energy $E_i \geq 0$, weight $w_i = g_i E_i$, and

$$\Phi = \begin{bmatrix} \phi_1^\top \\ \vdots \\ \phi_N^\top \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}, \quad W = \text{diag}(w_1, \dots, w_N).$$

Unified (weighted ridge) regression and residual.

$$\hat{\theta}_\lambda = \arg \min_{\theta \in \mathbb{R}^5} \|W^{1/2}(\Phi\theta - b)\|_2^2 + \lambda \|\theta\|_2^2, \quad \mathcal{L}_{\text{uni}, \lambda} = \frac{\|W^{1/2}(\Phi\hat{\theta}_\lambda - b)\|_2^2}{\text{tr}(W)}.$$

For $\lambda = 0$ write $\hat{\theta}$, $\mathcal{L}_{\text{uni}}$. For rotational/scaling/translational *slices*, restrict ϕ_i to the corresponding subspace to obtain $\mathcal{L}_{\text{uni}, \lambda}^{(\text{rot})}$, $\mathcal{L}_{\text{uni}, \lambda}^{(\text{scale})}$, $\mathcal{L}_{\text{uni}, \lambda}^{(\text{trans})}$.

Surrogate losses (as implemented).

- Translation: $\mathcal{L}_{\text{trans}} = \frac{\sum_i w_i e_i^2}{\sum_i w_i}$ with $e_i = \omega_t - q(\omega_x, \omega_y)$ for linear q .

- Rotation:

$$\mathcal{L}_{\text{rot}} = 1 - \frac{1}{2}(C_{\text{ring}} + C_{\text{rot}}),$$

where C_{rot} is a tilted-line energy ratio on (m, ω_t) and C_{ring} is the annular concentration.

- Scaling:

$$\mathcal{L}_{\text{scale}} = 1 - \frac{1}{2}(C_{\text{flow}} + S_{\text{trend}}).$$

Time windowing and interpolation. Let $h[t]$ be a temporal window (e.g., Hann) of length T ; its DFT is $\hat{h}(\omega_t)$. Temporal windowing induces convolution and *leakage*; denote the relative out-of-band energy by $\varepsilon_{\text{win}}(\Delta)$ (defined later). Polar/log-radius resampling via bilinear interpolation induces an error $\varepsilon_{\text{interp}}$.

Weights and gates. Weights are $w_i = g_i E_i$ with $g_i = g_E(i) \cdot g_{\text{obs}}(i)$. Assume $g_E(i) \in [g_{\min}, 1]$ (energy gate) and $g_{\text{obs}}(i) \in [\tilde{g}_{\min}, \tilde{g}_{\max}]$, hence

$$g_i \in [\underline{g}, \bar{g}], \quad \underline{g} := g_{\min} \tilde{g}_{\min} > 0, \quad \bar{g} := \tilde{g}_{\max} < \infty.$$

Narrow-band assumption (rotation/scaling). At each time, at least $(1 - \varepsilon)$ of the spatial spectral energy concentrates in a single annulus; $\varepsilon \in [0, 1)$.

Observability. Assume $\lambda_{\min}(\Phi^\top W \Phi) > 0$ (or $\lambda_{\min}(\Phi^\top W \Phi) + \lambda > 0$ with ridge) and the same for sliced subspaces.

B.2. Exactness for Ideal SIM(2)

Theorem B.1 (Asymptotic zero-residual). *Under a single SIM(2) motion, in the idealized continuous-time, infinite-window, and noise-free setting (no windowing or interpolation), the unified residual and slice surrogates tend to zero:*

$$\lim_{T \rightarrow \infty, \Delta \rightarrow 0^+} \left(\mathcal{L}_{\text{uni}}, \mathcal{L}_{\text{uni}}^{(\text{rot})}, \mathcal{L}_{\text{uni}}^{(\text{scale})}, \mathcal{L}_{\text{uni}}^{(\text{trans})}, \mathcal{L}_{\text{rot}}, \mathcal{L}_{\text{scale}}, \mathcal{L}_{\text{trans}} \right) = 0.$$

With finite windows and discrete sampling, these losses are lower bounded by window/leakage and interpolation terms characterized in Lemma B.3.

Sketch. Classical spectral support: translation on the plane $\omega_t + v_x \omega_x + v_y \omega_y = 0$; rotation on the line $\omega_t + \Omega m = 0$; scaling on the line $\omega_t + \alpha \nu = 0$. Least-squares projection to the true support yields zero residual; line-energy ratios equal 1; the annular entropy is 0. $\square$

B.3. Band-Capture and Window Leakage

Lemma B.2 (Weighted band-capture). *Let e_i be algebraic distances to a target line/plane, and define*

$$E_{\text{in}}(\Delta) = \sum_{|e_i| \leq \Delta} E_i, \quad E_{\text{all}} = \sum_i E_i.$$

If $w_i = g_i E_i$ with $g_i \in [\underline{g}, \bar{g}]$, then

$$1 - \frac{E_{\text{in}}(\Delta)}{E_{\text{all}}} \leq \frac{\bar{g}}{\underline{g}} \cdot \frac{1}{\Delta^2} \cdot \frac{\sum_i w_i e_i^2}{\sum_i w_i}. \quad (16)$$

Proof. Apply Chebyshev's inequality to the energy-weighted measure $p_i = E_i / E_{\text{all}}$: $\sum_{|e_i| > \Delta} p_i \leq \Delta^{-2} \sum_i p_i e_i^2$. Substitute $p_i = \frac{w_i}{g_i} / \sum_j \frac{w_j}{g_j}$ and bound g_i by $[\underline{g}, \bar{g}]$. $\square$

Lemma B.3 (Window leakage). *Let $\hat{h}(\omega_t)$ be the DFT of the temporal window $h[t]$. After convolution with $|H|^2$ in ω_t , the relative out-of-band energy outside $\pm \Delta$ satisfies*

$$\varepsilon_{\text{win}}(\Delta) \leq \frac{\sum_{|\omega_t| > \Delta} |\hat{h}(\omega_t)|^2}{\sum_{\omega_t} |\hat{h}(\omega_t)|^2}. \quad (17)$$

For Hann/Blackman, the RHS decreases monotonically in T and Δ .

B.4. Rotation: Surrogate Bounded by Unified Slice Residual

Rotation statistics. The tilted-line ratio on (m, ω_t) : $C_{\text{rot}} = \frac{E_{\text{in}}(\Delta)}{E_{\text{all}}}$ with $e_i = \omega_t + m \Omega^*$, and Ω^* given by energy-weighted LS:

$$\Omega^* = - \frac{\sum |\tilde{C}_m(\rho, \omega_t)|^2 \omega_t m}{\sum |\tilde{C}_m(\rho, \omega_t)|^2 m^2}.$$

Annular concentration: $C_{\text{ring}} = 1 - \frac{H}{\log N_r}$ where H is the entropy of the ring energy distribution.

Lemma B.4 (Annulus entropy bound). *If at least $(1 - \varepsilon)$ of the energy lies in one ring, then*

$$1 - C_{\text{ring}} = \frac{H}{\log N_r} \leq \frac{h(\varepsilon) + \varepsilon \log(N_r - 1)}{\log N_r}, \quad (18)$$

with $h(\varepsilon) = -\varepsilon \log \varepsilon - (1 - \varepsilon) \log(1 - \varepsilon)$.

Theorem B.5 (Rotation surrogate $\leq$ unified slice residual). *For $\Delta \geq 1$ (one temporal-frequency bin at least), with window/interp errors $\varepsilon_{\text{win}}(\Delta)$, $\varepsilon_{\text{interp}}$ and radial narrow-bandness ε , one has*

$$\mathcal{L}_{\text{rot}} \leq \frac{\bar{g}}{2g} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{uni}, \lambda=0}^{(\text{rot})} + \frac{h(\varepsilon) + \varepsilon \log(N_r - 1)}{2 \log N_r} + \varepsilon_{\text{win}}(\Delta) + \varepsilon_{\text{interp}}. \quad (19)$$

Proof. By Lemma B.2 with $e_i = \omega_t + m\Omega^*$ and Lemma B.3, $1 - C_{\text{rot}} \leq \frac{\bar{g}}{g} \Delta^{-2} \mathcal{L}_{\text{uni}}^{(\text{rot})} + \varepsilon_{\text{win}} + \varepsilon_{\text{interp}}$. By Lemma B.4, $1 - C_{\text{ring}} \leq \frac{h(\varepsilon) + \varepsilon \log(N_r - 1)}{\log N_r}$. Average the two bounds. $\square$

Remark (ridge). If a ridge version is used to estimate Ω^* , Theorem B.11 below implies $\mathcal{L}_{\text{uni}, \lambda}^{(\text{rot})} \leq \mathcal{L}_{\text{uni}, 0}^{(\text{rot})} + \frac{\lambda}{\text{tr}(W)} \|\theta_{\text{LS}}\|_2^2$; absorb the $O(\lambda)$ term on the RHS of (19).

B.5. Translation

Theorem B.6 (Band-in ratio vs. surrogate residual). *For any $\Delta \geq 1$, defining band-in energy around the fitted plane by $E_{\text{in}}(\Delta)$, one has*

$$1 - \frac{E_{\text{in}}(\Delta)}{E_{\text{all}}} \leq \frac{\bar{g}}{g} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{trans}} + \varepsilon_{\text{win}}(\Delta). \quad (20)$$

Here $e_i = \omega_{t,i} - q(\omega_{x,i}, \omega_{y,i})$ with the linear model $q(\omega_x, \omega_y) = \gamma_1 \omega_x + \gamma_2 \omega_y + \gamma_3$.

Proof. Lemma B.2 with $e_i = \omega_t - q(\omega_x, \omega_y)$, plus window broadening by Lemma B.3. $\square$

Theorem B.7 (Equivalence to unified slice residual). *Let S be the linear feature subspace for the weighted inner product $\langle u, v \rangle = \sum_i w_i u_i v_i$. There exist $c_1, c_2 > 0$ (depending on the weighted Gram matrix condition number) such that*

$$c_1 \mathcal{L}_{\text{uni}}^{(\text{trans})} \leq \mathcal{L}_{\text{trans}} \leq c_2 \mathcal{L}_{\text{uni}}^{(\text{trans})}. \quad (21)$$

Idea. Both are weighted least-squares projection residuals to the same subspace, up to reparametrization and vertical/orthogonal distance constants bounded by the subspace condition number. $\square$

B.6. Scaling: Surrogate Bounded by Unified Slice Residual

Shift model on log-radius–time. On discrete (i, t) , assume

$$E(i, t) = A(i - u(t)) + \eta(i, t), \quad (22)$$

with A unimodal Lipschitz, $u(t)$ monotone C^1 , and perturbation η .

Lemma B.8 (Gradient alignment). *Let $\nabla_r E_{i,t} = E_{i+1,t} - E_{i,t}$, $\nabla_t E_{i,t} = E_{i,t+1} - E_{i,t}$, and normalized fields $\hat{\nabla}_r = \nabla_r E / \|\nabla_r E\|_2$, $\hat{\nabla}_t = \nabla_t E / \|\nabla_t E\|_2$. If $\eta = 0$, then $\nabla_t E = -u'(t) \nabla_r E$ so $C_{\text{flow}} = |\langle \hat{\nabla}_r, \hat{\nabla}_t \rangle| = 1$. If $\|\eta\|_2 \leq \varepsilon \|\nabla_r E\|_2$, then*

$$C_{\text{flow}} \geq \frac{|\bar{u}'| - c\varepsilon}{|\bar{u}'| + c\varepsilon}, \quad (23)$$

where $\bar{u}'$ is the window-average of u' and $c > 0$ depends on discrete derivative constants.

Lemma B.9 (Centroid trend). *Let $\rho_c(t) = \sum_i i E(i, t) / \sum_i E(i, t)$. If $\eta = 0$ and A is unimodal, $\rho_c(t)$ is monotone with $u(t)$, and $|\text{corr}(r_c, t)| \rightarrow 1$ when $u(t)$ is near-linear. With perturbation $\|\eta\|_2 \leq \varepsilon \|A\|_2$, one has $|\text{corr}(r_c, t)| \geq 1 - \delta(\varepsilon)$, hence*

$$S_{\text{trend}} = |\text{corr}(\rho_c, t)| \geq 1 - \delta(\varepsilon). \quad (24)$$

Theorem B.10 (Scaling surrogate $\leq$ unified slice residual). *Let $C_{\text{scale}} = E_{\text{in}}(\Delta)/E_{\text{all}}$ be defined on (ν, ω_t) with $e_i = \omega_t + \alpha^* \nu$ and α^* from weighted LS. Then*

$$1 - C_{\text{scale}} \leq \frac{\bar{g}}{\underline{g}} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{uni}, \lambda=0}^{(\text{scale})} + \varepsilon_{\text{win}}(\Delta) + \varepsilon_{\text{interp}}. \quad (25)$$

If in addition (22) holds and Lemmas B.8–B.9 apply, there exists $\delta_{\text{flow}} \geq 0$ s.t.

$$\mathcal{L}_{\text{scale}} \leq \frac{\bar{g}}{2\underline{g}} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{uni}, \lambda=0}^{(\text{scale})} + \varepsilon_{\text{win}}(\Delta) + \varepsilon_{\text{interp}} + \frac{1}{2} \delta_{\text{flow}}. \quad (26)$$

Proof. Equation (25) follows from Lemmas B.2–B.3. Moreover, $\frac{1}{2}(C_{\text{flow}} + S_{\text{trend}}) \geq C_{\text{scale}} - \delta_{\text{flow}}/2$ by Lemmas B.8–B.9. Since $\mathcal{L}_{\text{scale}} = 1 - \frac{1}{2}(C_{\text{flow}} + S_{\text{trend}})$, we obtain (26). $\square$

B.7. Ridge Regression: Residual and Consistency

Theorem B.11 (Ridge residual and consistency). *Let $X = W^{1/2}\Phi$, $y = W^{1/2}b$, and*

$$\hat{\theta}_\lambda = \arg \min_{\theta} \|X\theta - y\|_2^2 + \lambda \|\theta\|_2^2.$$

Let θ_{LS} be the LS solution and $r_\star = \|X\theta_{\text{LS}} - y\|_2^2$. Then

$$\|X\hat{\theta}_\lambda - y\|_2^2 \leq r_\star + \lambda \|\theta_{\text{LS}}\|_2^2. \quad (27)$$

If $y = X\theta^\star + \xi$, $\mathbb{E}[\xi] = 0$, $\text{Cov}(\xi) = \sigma^2 I$, and $\lambda = \lambda_N \rightarrow 0$, $N\lambda_N \rightarrow \infty$, then $\hat{\theta}_\lambda \xrightarrow{P} \theta^\star$ and $\mathbb{E}[\mathcal{L}_{\text{uni}, \lambda}] \rightarrow \sigma^2 c$ for some constant c .

Proof. Evaluate the ridge objective at $\hat{\theta}_\lambda$ vs. θ_{LS} to get (27). Consistency follows from standard ridge regression results with weights absorbed into X . $\square$

Corollary (ridge absorption). In Theorems B.5 and B.10, one may replace $\mathcal{L}_{\text{uni}, 0}^{(\cdot)}$ by $\mathcal{L}_{\text{uni}, \lambda}^{(\cdot)}$ at the cost of an additive $O(\lambda)$ term.

B.8. Consolidated Bounds

Under the stated assumptions ($w_i = g_i E_i$, $g_i \in [\underline{g}, \bar{g}]$, $\Delta \geq 1$, window/interp bounded, narrow-band rotation/scaling), the three surrogate losses satisfy

$$\begin{aligned} \mathcal{L}_{\text{rot}} &\leq \frac{\bar{g}}{2\underline{g}} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{uni}, \lambda}^{(\text{rot})} + \frac{h(\varepsilon) + \varepsilon \log(N_r - 1)}{2 \log N_r} + \varepsilon_{\text{win}}(\Delta) + \varepsilon_{\text{interp}} + O(\lambda), \\ \mathcal{L}_{\text{scale}} &\leq \frac{\bar{g}}{2\underline{g}} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{uni}, \lambda}^{(\text{scale})} + \varepsilon_{\text{win}}(\Delta) + \varepsilon_{\text{interp}} + \frac{1}{2} \delta_{\text{flow}} + O(\lambda), \\ \mathcal{L}_{\text{trans}} &\leq \frac{\bar{g}}{\underline{g}} \cdot \frac{1}{\Delta^2} \mathcal{L}_{\text{uni}, \lambda}^{(\text{trans})} + \varepsilon_{\text{win}}(\Delta) + O(\lambda). \end{aligned} \quad (28)$$

C. Spectral low-pass truncation: model, bounds, and sanity check

Low-pass truncation. We apply a per-dimension low-frequency cutoff $\varrho=0.3$ to the 3D DFT lattice of $(\omega_t, \omega_x, \omega_y)$. This defines a low-frequency cube

$$C_\infty(\varrho) = \{\|\hat{\omega}\|_\infty \leq \varrho\},$$

under the normalized frequency convention used below. The same cutoff is used across all main experiments.

Spectral model and ball-retained energy. Natural video spectra are well-approximated by a radial power law (Ruderman & Bialek, 1994; Dong & Atick, 1995):

$$E(\omega_x, \omega_y, \omega_t) \propto (\omega_x^2 + \omega_y^2 + \omega_t^2)^{-\kappa}, \quad \kappa \approx 1.8.$$

For analysis, we pass to dimensionless frequencies $\hat{\omega}_t = \omega_t/(T-1)$, $\hat{\omega}_x = \omega_x/(W-1)$, $\hat{\omega}_y = \omega_y/(H-1)$, and define the radial frequency

$$r = \sqrt{\hat{\omega}_t^2 + \hat{\omega}_x^2 + \hat{\omega}_y^2}.$$

Let $\varepsilon > 0$ be the minimum nonzero radius on the discrete grid and let

$$R = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}.$$

For a relative radial cutoff $a \in (0, 1]$, the ball of radius aR retains

$$E_{\text{tot}} \propto \int_{\varepsilon}^R 4\pi r^{2-2\kappa} dr = \frac{4\pi}{3-2\kappa} (R^{3-2\kappa} - \varepsilon^{3-2\kappa}), \quad (29)$$

$$E_{\text{ball}}(a) \propto \int_{\varepsilon}^{aR} 4\pi r^{2-2\kappa} dr = \frac{4\pi}{3-2\kappa} ((aR)^{3-2\kappa} - \varepsilon^{3-2\kappa}). \quad (30)$$

Hence the retained-energy fraction for the ball is

$$\eta_{\text{ball}}(a) = \frac{E_{\text{ball}}(a)}{E_{\text{tot}}} = \frac{(aR)^{3-2\kappa} - \varepsilon^{3-2\kappa}}{R^{3-2\kappa} - \varepsilon^{3-2\kappa}}. \quad (31)$$

For natural videos with $\kappa > 1.5$, let $\beta = 2\kappa - 3 > 0$. Then

$$\eta_{\text{ball}}(a) = 1 - \frac{a^{-\beta} - 1}{(R/\varepsilon)^{\beta} - 1}. \quad (32)$$

Cube vs. ball with normalization-aware bounds. Our implementation uses a cube rather than a ball. Let

$$B_2(r) = \{\|\hat{\omega}\|_2 \leq r\}$$

and let $E(S)$ denote the spectral energy summed or integrated over a frequency set S . The actual retained fraction of the cube truncation is

$$\eta_{\text{cube}}(\varrho) = \frac{E(C_{\infty}(\varrho))}{E(C_{\infty}(1))}.$$

The coordinate-radius inclusions are

$$B_2(\varrho) \subset C_{\infty}(\varrho) \subset B_2(\sqrt{3}\varrho), \quad B_2(1) \subset C_{\infty}(1) \subset B_2(\sqrt{3}).$$

Since $\eta_{\text{ball}}(a)$ in Eq. (31) is parameterized by the relative radial cutoff $a = r/R$ with $R = \sqrt{3}$, these inclusions imply the conservative normalization-aware bound

$$\eta_{\text{ball}}(\varrho/\sqrt{3}) \leq \eta_{\text{cube}}(\varrho) \leq \min \left\{ 1, \frac{\eta_{\text{ball}}(\varrho)}{\eta_{\text{ball}}(1/\sqrt{3})} \right\}. \quad (33)$$

Numerical sanity check. With $\kappa=1.8$ and $\varrho=0.3$, typical video sizes such as $T=16$, $H=W=224$ give $R/\varepsilon \approx 3.87 \times 10^2$. Using Eq. (33), we obtain approximately

$$0.946 \leq \eta_{\text{cube}}(0.3) \leq 0.981.$$

Empirically on 1k random videos we measure 97.5% retained spectral energy, which lies within this conservative range.

Table 9. Fixed values for main experiments.

Item	Value
low-pass ratio	0.3
Energy threshold τ_E / smoothing factor f	0.10 / 10
Rings N_r / angular bins M / log-radius N_ξ	20 / 24 / 24
Soft-ring edge sharpness	20
Tilted-line tolerance Δ	1 (temporal-frequency bin)
Ridge λ / numeric ε	10^{-3} / 10^{-8}
Softmax temperature τ	0.1
Physics-loss mixing weight	0.1
Precision policy	Spectral/solvers FP32; others BF16
Temporal window	Hann

D. Implementation Details

D.1. Translational-loss WLS details

We instantiate the WLS fitting as follows. For the compact relation $A\beta_{\text{tr}} - b = 0$, the per-sample row and vectors are:

- **Translation (plane):** $A \in \mathbb{R}^{N \times 3}$ with rows $A_i = (\omega_{x,i}, \omega_{y,i}, 1)$, $\beta_{\text{tr}} = [v_x, v_y, b_0]^\top$, and $b_i = -\omega_{t,i}$.

Weights follow the energy/observability gate in App. A.6. The normalized residual is

$$\mathcal{L}_{\text{trans}} = \frac{\sum_i \mathbf{W}_{ii} (A_i \hat{\beta}_{\text{tr}} - b_i)^2}{\sum_i \mathbf{W}_{ii}}.$$

Implementation details (regularization, precision policy) follow the general solver notes in App. §A.6.

D.2. Rotational-loss details

Ring concentration. We partition the spatial frequency plane into N_r concentric annuli with masks $\{\mathcal{M}_i\}_{i=1}^{N_r}$ and define

$$E_k(t) = \frac{\sum_{(\omega_x, \omega_y) \in \mathcal{M}_k} E(\omega_x, \omega_y, t)}{\sum_{j=1}^{N_r} \sum_{(\omega_x, \omega_y) \in \mathcal{M}_j} E(\omega_x, \omega_y, t) + \epsilon_{\text{stab}}}, \quad H_{\text{ring}}(t) = -\sum_{k=1}^{N_r} E_k(t) \log(E_k(t) + \epsilon_{\text{stab}}), \quad \bar{H}_{\text{ring}} = \frac{1}{T} \sum_{t=1}^T H_{\text{ring}}(t), \quad (34)$$

where $E(\omega_x, \omega_y, t) = |\hat{V}(\omega_x, \omega_y, t)|^2$ is the spatiotemporal spectral energy, and $\epsilon > 0$ ensures numerical stability. We set $C_{\text{ring}} = 1 - \frac{\bar{H}_{\text{ring}}}{\log N_r}$.

Tilted-line energy along $\omega_t + \Omega m = 0$. We resample the spectrum to polar coordinates (ρ, θ, t) via bilinear interpolation $\hat{V}(\rho, \theta, t)$, then take the angular DFT and temporal DFT:

$$C_m(\rho, t) = \frac{1}{2\pi} \int_0^{2\pi} \hat{V}(\rho, \theta, t) e^{-im\theta} d\theta, \quad \tilde{C}_m(\rho, \omega_t) = \text{DFT}_t\{C_m(\rho, t)\}. \quad (35)$$

An energy-weighted least squares gives the angular velocity estimate

$$\Omega^\star = -\frac{\sum_\rho \sum_{m \neq 0} \sum_{\omega_t} |\tilde{C}_m(\rho, \omega_t)|^2 \omega_t m}{\sum_\rho \sum_{m \neq 0} \sum_{\omega_t} |\tilde{C}_m(\rho, \omega_t)|^2 m^2}. \quad (36)$$

We measure how much energy lies within a narrow band of width Δ around the ideal line $\omega_t + m\Omega^\star = 0$:

$$E_{\text{line}} = \sum_\rho \sum_{m \neq 0} \sum_{|\omega_t + m\Omega^\star| \leq \Delta} |\tilde{C}_m(\rho, \omega_t)|^2, \quad E_{\text{all}} = \sum_\rho \sum_{m \neq 0} \sum_{\omega_t} |\tilde{C}_m(\rho, \omega_t)|^2, \quad (37)$$

and define the tilted-line energy ratio

$$C_{\text{rot}} = \frac{E_{\text{line}}}{E_{\text{all}}}. \quad (38)$$

By default Δ is one temporal-frequency bin and $m = 0$ is excluded. Optional observability weights that downweight tiny $|m|$ or low-energy bins can be absorbed into the sums. Implementation options (polar LUT, windowing, precision, caching) are detailed in App. A.6.

D.3. Robustness to Brightness and Rendering Perturbations

To substantiate the robustness benefit of the frequency-domain diagnostics, we test the three motion losses under controlled brightness, blur, noise, and JPEG perturbations. Across translation, rotation, and scaling, the matched loss changes only slightly under these perturbations, and the selected motion type remains correct in every case.

Table 10. Robustness of the motion diagnostics to brightness and rendering perturbations. The matched loss changes only slightly, and the selected motion type remains correct in every case.

Perturbation	Translation sequence		Rotation sequence		Scaling sequence	
	$\mathcal{L}_{\text{trans}}$	Argmin	$\mathcal{L}_{\text{rot}}$	Argmin	$\mathcal{L}_{\text{scale}}$	Argmin
Clean	0.134	Translation	0.336	Rotation	0.185	Scaling
Brightness $\times 1.3$	0.135	Translation	0.336	Rotation	0.188	Scaling
Brightness $+0.1$	0.135	Translation	0.340	Rotation	0.188	Scaling
Blur ($k = 5$)	0.140	Translation	0.336	Rotation	0.187	Scaling
Noise ($\sigma = 0.03$)	0.135	Translation	0.338	Rotation	0.185	Scaling
JPEG ($q = 30$)	0.135	Translation	0.336	Rotation	0.186	Scaling

These results provide a targeted robustness check suggesting that the frequency-domain diagnostics are relatively tolerant to small brightness and rendering perturbations.

D.4. Sensitivity to Weighting Temperature and Low-Pass Ratio

We evaluate the sensitivity of our method to the weighting temperature τ and low-pass ratio ϱ on Wan 2.1-14B. As shown in Table 11, all tested variants improve Visual Quality over the no-motion-loss baseline, and the default setting achieves the best overall balance.

Table 11. Sensitivity to the weighting temperature τ and low-pass ratio ϱ on Wan 2.1-14B.

Setting	VQA_A $\uparrow$	VQA_T $\uparrow$	IS $\uparrow$	Vis. Qual. $\uparrow$
No motion loss	67.48	64.43	16.28	54.18
$\tau = 0.05, \varrho = 0.3$	69.61	64.97	17.24	61.95
$\tau = 0.1, \varrho = 0.3$ (default)	71.18	66.98	18.09	62.65
$\tau = 0.2, \varrho = 0.3$	69.51	65.03	17.58	60.69
$\tau = 0.1, \varrho = 0.2$	68.26	66.91	17.73	61.93
$\tau = 0.1, \varrho = 0.4$	69.05	67.97	17.55	62.06

D.5. Timestep Weighting and Temporal Window

We ablate several strategies for weighting the physics loss across diffusion timesteps. Uniform weighting applies the loss equally at all timesteps, while low-noise-only weighting applies the loss only when $t/T_{\text{diff}} \leq 0.5$. We also compare the default linear decay $w = 1 - t/T_{\text{diff}}$ with an SNR-based weighting strategy. **Predicted-clean estimate.** The auxiliary frequency loss is applied to the predicted clean sample $\hat{x}_0$, i.e., the standard clean-sample reconstruction from the diffusion model prediction. This estimate can be viewed as a posterior-mean/Tweedie-style estimate, and becomes overly smooth and less reliable at high noise levels. We therefore compute the loss only on the low-frequency spectrum and use timestep weighting to reduce the contribution of high-noise steps.

Table 12. Ablation of timestep weighting strategies. Uniform weighting applies the physics loss even at high-noise timesteps and substantially degrades performance. The default linear decay provides the best overall balance.

Strategy	VQA_A $\uparrow$	Visual Quality $\uparrow$
Uniform weight	56.32	59.06
Low-noise only ($t/T_{\text{diff}} \leq 0.5$)	69.17	62.35
Linear decay, default	71.18	62.65
SNR-based	70.44	62.19

These results show that applying the physics loss uniformly across all noise levels is harmful, consistent with the predicted clean sample $\hat{x}_0$ being less reliable at high noise levels. Timestep-weighted variants preserve the useful low-frequency motion signal, with linear decay performing best in our experiments.

To verify that this trend is not specific to a single backbone, we repeat the same timestep-weighting ablation on MVDIT using identical hyperparameters. As shown in Table 13, the same qualitative pattern holds: uniform weighting performs worst, while the default linear decay achieves the best overall balance.

Table 13. Cross-backbone ablation of timestep weighting strategies on MVDIT. The same qualitative trend as Table 12 is observed: applying the physics loss uniformly across all timesteps degrades performance, while the default linear decay gives the best overall result.

Strategy	VQA_A $\uparrow$	Visual Quality $\uparrow$
Uniform weight	59.85	56.79
Low-noise only ($t/T_{\text{diff}} \leq 0.5$)	67.30	61.66
Linear decay, default	69.30	62.43
SNR-based	67.95	61.73

Together with the Wan 2.1-14B ablation, this suggests that linear decay is a simple and stable default across different diffusion backbones, rather than a setting tuned to a particular noise schedule or architecture.

We use a temporal FFT window of $K = 16$ frames. To assess sensitivity, we evaluate $\mathcal{L}_{\text{trans}}$ on synthetic sequences with different apparent velocities, where v_x denotes horizontal displacement per frame normalized by frame width.

Table 14. Sensitivity of $\mathcal{L}_{\text{trans}}$ to temporal window length. Larger windows better capture slow motions, while shorter windows are more suitable for fast motions. We use $K = 16$ as a compromise.

Speed group	$K = 12$	$K = 16$	$K = 20$
Slow ($v_x \leq 0.01$)	0.168	0.099	0.082
Medium ($0.01 < v_x \leq 0.02$)	0.047	0.046	0.115
Fast ($v_x > 0.02$)	0.066	0.098	0.144

The results show the expected trade-off: larger windows help slow motions but are less suitable for fast motions. We use $K = 16$ as a fixed default because it provides a reasonable compromise across speed groups. For reference, on the 605 SIM(2)-approximable OpenVID-1M clips, a clip-level normalized apparent-displacement proxy gives median / p90 / p95 values of 1.34% / 3.22% / 4.05% frame-widths per frame, suggesting that the synthetic sweep covers common motions and much of the upper range in this subset. Adaptive data-driven selection of K , or a multi-window variant combining several temporal scales, is a natural extension for datasets with substantially different motion statistics.

E. LLM Usage

Large language models (LLMs) were used only as tools and are not authors. Specifically: (1) we used an LLM to assist with prompt stratification into “simple” vs. “complex” motion following a rubric defined by us; and (2) we used an LLM for grammar and style polishing.