# DOES SEMANTIC NOISE INITIALIZATION TRANSFER FROM IMAGES TO VIDEOS? A PAIRED DIAGNOSTIC STUDY

**Yixiao Jing**[1,*] **Chaoyu Zhang**[1,*] **Zixuan Zhong**[2,*] **Peizhou Huang**[1,†]
[1]University of Michigan [2]University College London
peizhou@umich.edu

## ABSTRACT

Semantic noise initialization has been reported to improve robustness and controllability in image diffusion models. Whether these gains transfer to text-to-video (T2V) generation remains unclear, since temporal coupling can introduce extra degrees of freedom and instability. We benchmark semantic noise initialization against standard Gaussian noise using a frozen VideoCrafter-style T2V diffusion backbone and VBench on 100 prompts. Using prompt-level paired tests with bootstrap confidence intervals and a sign-flip permutation test, we observe a small positive trend on temporal-related dimensions; however, the 95% confidence interval includes zero ($p \approx 0.17$) and the overall score remains on par with the baseline. To understand this outcome, we analyze the induced perturbations in noise space and find patterns consistent with weak or unstable signal. We recommend prompt-level paired evaluation and noise space diagnostics as standard practice when studying initialization schemes for T2V diffusion. Our code and evaluation scripts are available at: https://github.com/klkds/golden-noise-transfer

## 1 INTRODUCTION

Text-to-video (T2V) diffusion models are sensitive to random seeds: different initial Gaussian noises can yield large semantic and motion variations under the same prompt, complicating controllability and reliable comparisonSingh et al. (2025). Recent image-generation work suggests that teacher-aligned noise initialization can improve robustness by moving the starting noise distribution closer to a teacher's preferred region in noise spaceAhn et al. (2024); Li et al. (2025); Eyring et al. (2025). A natural hypothesis is that videos may benefit even more, since temporal dynamics amplify seed-induced varianceHo et al. (2022).

In this work, we conduct a focused diagnostic study on transferring semantic noise initialization from image generation to video diffusion. We instantiate a lightweight noise mapper (named NPNet) on top of a frozen video diffusion backbone (VideoCrafter-style) and evaluate on 100 prompts using VBench. Crucially, we report the paired statistical testing over prompts (bootstrap CI and permutation test), which is necessary when effect sizes are small relative to prompt-level varianceEfron (1992); Dror et al. (2018). We further analyze the underlying mechanisms that govern how semantic ("golden") noise perturbations propagate in the spatio-temporal diffusion process, by conducting a cross-model diagnostic on Open-Sora2 and VideoCrafter, which employ different video diffusion sampling mechanisms.

**Key findings.** (i) On VideoCrafter, semantic noise mapping shows a slight but statistically insignificant positive trend on temporal metrics ($p \approx 0.17$), with overall scores remaining on par with the baseline. (ii) Prompt-level variance dominates the effect size, placing this approach in a low signal-to-noise ratio (low-SNR) regime. (iii) Noise-space diagnostics indicate that while induced semantic perturbations are structured, their directional stability and spatiotemporal frequency profiles vary substantially between Open-Sora2 and VideoCrafter, helping explain the inconsistent temporal gains across models.

**Contributions.**

- Reproducible paired evaluation of semantic noise initialization on a VideoCrafter-style T2V diffusion model over 100 prompts.
- Prompt-level significance testing (bootstrap CI and permutation test), clarifying that temporal-metric trends are not statistically reliable under this setting.
- We develop cross-model noise-space diagnostics that characterize the directional stability and spatiotemporal frequency structure of semantic perturbations, enabling systematic comparison across video diffusion backbones.

## 2 Related Work

**Diffusion-based T2V Generation.** Diffusion-based generators are a dominant paradigm for high-fidelity image and video synthesis. Modern T2V diffusion largely inherits DDPM and latent-diffusion formulations, injecting text conditioning via cross-attention and classifier-free guidance (CFG) Ho et al. (2020); Rombach et al. (2022); Ho & Salimans (2022). Recent systems (e.g., VideoCrafter-style backbones) extend denoising to spatio-temporal latents to model motion and temporal coherence Chen et al. (2023), but typically keep the highest-noise initialization as an isotropic Gaussian draw, leaving *initialization* relatively underexplored.

**Seed Sensitivity and Noise-space Learning.** T2V diffusion is sensitive to random seeds, motivating noise-space transformations on frozen backbones, exemplified by Golden Noise: semantically aligned noise targets are constructed with a teacher diffusion model and a lightweight mapper is trained from standard Gaussian noise to the learned distribution Zhou et al. (2025). Inversion methods (commonly DDIM inversion) provide practical trajectories between noise and denoised latents for target construction Song et al. (2020). Our semantic noise initialization follows this teacher-in-noise principle and adapts it to spatio-temporal video latents to reduce seed-induced variance without retraining the full backbone.

**Evaluation of T2V Quality.** T2V evaluation typically combines semantic alignment, which is often CLIP-based Radford et al. (2021), perceptual similarity and diversity (e.g., Learned Perceptual Image Patch Similarity, LPIPS) Zhang et al. (2018), and other temporal consistency metrics; suites on benchmarks such as VBench offer standardized protocols Huang et al. (2024). To isolate the impact of initialization, we keep the backbone, sampler, prompt, and seed fixed, and make changes only the initial noise so that differences in temporal behavior can be attributed to the learned initializer.

## 3 Methodology

**Problem Setup.** Let $G_\theta$ denote a frozen T2V diffusion generator that samples a video from an initial noise latent $z_T \sim \mathcal{N}(0, I)$ and a prompt $p$. Seed sensitivity arises because perturbations in $z_T$ can induce variations in the generated spatio-temporal latent trajectory.

**Semantic (Golden) Noise Targets.** Following the teacher-in-noise principle, we consider a semantic target noise $z_T^\star$ (golden noise) that is more aligned with the prompt-conditioned denoising trajectory. In practice, $z_T^\star$ can be obtained by optimization or inversion-style procedures that search in noise space for an initialization yielding higher semantic and temporal quality under a fixed backbone and sampler. This extraction is computationally non-trivial for videos, as evaluating a candidate noise typically requires running the spatio-temporal denoising process.

**NPNet: A Lightweight Noise Mapper.** We train a lightweight mapper $f_\phi$ that transforms standard Gaussian noise into a semantic initialization:

$$\hat{z}_T = f_\phi(z_T, p), \tag{1}$$

where $f_\phi$ is conditioned on the prompt $p$ (e.g., via text embedding injection). We train $f_\phi$ to approximate the extracted targets using a regression loss:

$$\mathcal{L}(\phi) = \mathbb{E}\left[\|f_\phi(z_T, p) - z_T^\star\|_2^2\right], \tag{2}$$

while keeping $G_\theta$ frozen. At inference, we replace $z_T$ with $\hat{z}_T$ and sample with the same backbone, sampler, and guidance.

**Cost Model and Practical Overhead.** Unlike images, extracting $z_T^\star$ for videos can incur substantial overhead because each candidate requires spatio-temporal denoising evaluation. This motivates reporting the cost–benefit trade-off explicitly: even when metric gains exist, they must outweigh the additional compute for target extraction and training.

# 4 EXPERIMENTS

## 4.1 SETUP

Unless otherwise stated, we evaluate on 100 prompts sampled from the VBench prompt set with 5 random seeds per prompt. For each prompt–seed pair, we keep the text prompt, the backbone $G_\theta$, the scheduler, and the CFG configuration identical, and change only the initial latent at the highest-noise timestep ($t = T$). The baseline samples $z_T \sim \mathcal{N}(0, I)$, while NPNet deterministically maps the same sampled noise using the same prompt embedding, $\hat{z}_T = f_\phi(z_T, p)$, and then runs the same sampling procedure thereafter. Thus, both methods share the same conditioning in the diffusion backbone; the only difference is the initialization at $t = T$ (i.e., the initial latent fed into the sampler). For statistical analysis, we use prompt-level paired comparisons: we first average each metric over the 5 seeds for each prompt, and then compute paired differences across the 100 prompts (i.e., the statistical unit is the prompt, $N = 100$), unless explicitly noted otherwise.

## 4.2 QUANTITATIVE EVALUATION

We adopt VBench and use temporal_style as the primary temporal metric, as it most directly captures temporal dynamics like flicker and jitter, and exhibits lower prompt-level noise than composite temporal aggregates in our setting. To assess reliability across prompts, we perform prompt-level paired tests over the 100 prompts. Concretely, for each prompt we average the metric over the 5 seeds for the baseline and for NPNet, then form a paired difference (NPNet minus baseline). We report (i) a bootstrap confidence interval (CI) for the mean paired difference and (ii) a paired sign-flip permutation test for the null hypothesis of zero mean difference.

Table 1: Mean VBench scores on 100 prompts.

| Dimension | Baseline | NPNet | $\Delta$ |
|---|---|---|---|
| aesthetic_quality | 0.638083 | 0.634992 | -0.003091 |
| imaging_quality | 0.715084 | 0.707932 | -0.007151 |
| background_consistency | 0.976592 | 0.976812 | +0.000220 |
| subject_consistency | 0.977814 | 0.978034 | +0.000220 |
| temporal_style | 0.076961 | 0.078716 | +0.001754 |

Table 1 reports seed-averaged mean scores over prompts for reference. All statistical claims are based on the prompt-level paired testing with prompt as the unit ($N = 100$; shown in Appendix E, Table 4); the improvement on temporal_style is not significant (95% CI crosses zero; $p = 0.1687$). Although aggregate metrics are neutral and improvements are not statistically significant, we occasionally observe prompt-specific gains in fine textures (e.g., fur/scales); see Appendix A.

## 4.3 QUALITATIVE NOISE-SPACE DIAGNOSTICS

To interpret the quantitative trade-off, we analyze the geometry and spatiotemporal frequency characteristics of golden noise $z_g$ relative to standard Gaussian noise $z$. Beyond characterizing VideoCrafter alone, we include a cross-model diagnostic with Open-Sora2 to assess whether the induced noise structure is intrinsic or dependent on the sampling dynamics. We define the displacement $d = z_g - z$ and aggregate statistics over 100 prompts $\times$ 5 seeds. Formal definitions of the geometry- and frequency-based metrics are provided in Appendix G.

Table 2: Global geometry and directional consistency of golden-induced perturbations for Open-Sora2 and VideoCrafter.

| Metric | Open-Sora2 | VideoCrafter |
|---|---|---|
| $\|z_g - z\|/\|z\|$ | 0.022 | 0.110 |
| $\cos(z, z_g)$ | 0.9997 | 0.9939 |
| Directional Stability (DirStab) | 0.631 | 0.200 |
| $\mathrm{CV}_{\|d\|}$ | 0.064 | 0.110 |
| Explained Variance Ratio (EVR1) | 0.464 | 0.343 |

Table 3: Frequency summary: global invariance between $z$ and $z_g$, and spatiotemporal characteristics of the displacement $d = z_g - z$.

| | Open-Sora2 | | | VideoCrafter | | |
|---|---|---|---|---|---|---|
| Metric | Mean | P10 | P90 | Mean | P10 | P90 |
| $\Delta$ Spatial HF ($z_g - z$) | -0.00049 | -0.00114 | 0.00006 | -0.01489 | -0.02086 | -0.00745 |
| $\mathrm{sp\_hf}(d)$ | 0.17996 | 0.06667 | 0.27212 | 0.24804 | 0.14243 | 0.42861 |
| $\mathrm{t\_hf}(d)$ | 0.85502 | 0.82826 | 0.87007 | 0.59898 | 0.45872 | 0.73720 |
| $\mathrm{tDiffRel}(d)$ | 0.60583 | 0.41060 | 0.75598 | 0.54638 | 0.45229 | 0.64628 |

Table 2 shows that $z_g$ stays close to $z$ in both models, but the displacement $d = z_g - z$ is considerably more aligned across seeds in Open-Sora2 than in VideoCrafter (DirStab/EVR1). Table 3 indicates near-invariant input frequency in Open-Sora2 ($\Delta$ Spatial HF $\approx 0$) but a systematic frequency shift in VideoCrafter, while the displacement exhibits model-dependent spatiotemporal structure.

## 5 Discussion

Our paired evaluation indicates that semantic noise initialization yields at most a small positive trend on temporal-related VBench dimensions, but the effect is not statistically significant under 100 prompts. The qualitative diagnostics help interpret this low-SNR outcome: in Open-Sora2, golden noise remains extremely close to the Gaussian prior in global geometry, yet induces a structured, prompt-conditioned displacement that is consistent across seeds (DirStab/EVR1). Crucially, the change concentrates in the displacement $d = z_g - z$ rather than the global spectrum of $z_g$, and exhibits a spatiotemporal imbalance: spatially smooth but temporally high-frequency.

In contrast, when evaluated on VideoCrafter with DDIM sampling, the induced displacement becomes substantially more dispersed in direction across seeds. As a consequence, the temporal high-frequency components are less concentrated and their relative dominance is reduced, resulting in a weaker amplification of temporal instability. We attribute this effect to the path-dependent dynamics of DDIM, which tend to rotate and diffuse initial directional perturbations across sampling steps.

This offers a coherent mechanism for the observed trade-off: a stable low-frequency bias may support coarse temporal coherence, while temporally jittery components can amplify flicker/jitter through temporal coupling during denoising, degrading perceptual quality. Taken together, these findings suggest that directly transferring image-style semantic/golden initialization to videos can enter a regime where the signal exists and is structured, but its temporal frequency characteristics make the net gain fragile under standard benchmark protocols.

## 6 Conclusion

We studied whether semantic (teacher-aligned) noise initialization transfers from images to T2V diffusion using a frozen video generation model's backbone and a lightweight prompt-conditioned mapper (NPNet). On 100 prompts with prompt-level paired tests, NPNet shows a small positive trend on temporal-related metrics but no statistically significant improvement, leaving overall quality near parity with the Gaussian baseline. Our noise-space diagnostics suggest the method mainly adds a structured spatiotemporal displacement rather than changing the global spectrum, which can yield fragile temporal gains and occasional perceptual degradation.

## REFERENCES

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024.

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

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, 2018.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Luca Eyring, Shyamgopal Karthik, Alexey Dosovitskiy, Nataniel Ruiz, and Zeynep Akata. Noise hypernetworks: Amortizing test-time compute in diffusion models. *arXiv preprint arXiv:2508.09968*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Zeming Li, Xiangyue Liu, Xiangyu Zhang, Ping Tan, and Heung-Yeung Shum. Noisear: Autoregressing initial noise prior for diffusion models. *arXiv preprint arXiv:2506.01337*, 2025.

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

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

Gurprit Singh, Xingchang Huang, Jente Vandersanden, Cengiz Oztireli, and Niloy Mitra. Demystifying noise: Role of randomness in generative ai: Demystifying randomness in generative ai. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Courses*, pp. 1–7, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17688–17697, 2025.

## A Additional Qualitative Results

In this section, we provide enlarged visual comparisons to supplement the quantitative results discussed in Section 4. Figure 1 contrasts the standard Gaussian initialization (Baseline) with our proposed semantic noise initialization (NPNet) across three distinct prompts. As observed, our method yields sharper high-frequency details and more consistent textures, particularly in challenging regions such as animal fur and skin scales.


Figure 1: **Qualitative comparison on VideoCrafter.** (Full size view) We visualize samples from Baseline (columns 1, 3, 5) versus our NPNet initialization (columns 2, 4, 6). Our method improves visual fidelity and detail consistency (e.g., the fur of the squirrel and the scales of the lizard) without changing the diffusion backbone.

## B Limitations

First, our conclusions are scoped to a VideoCrafter-style backbone and a fixed sampling/guidance configuration; different backbones, samplers, or guidance scales may shift the balance between temporal coherence and perceptual quality. Second, our evaluation relies on VBench metrics and prompt-level paired testing; while this is appropriate for small effects, it may not fully capture human preference or prompt-specific failure modes (e.g., rare motion artifacts). Third, the qualitative analysis focuses on aggregate noise-space statistics and frequency summaries; these diagnostics explain correlations but do not constitute a causal proof of how specific frequency components propagate through the denoising dynamics. Finally, extracting or optimizing golden targets for video remains non-trivial in compute; thus, even when improvements exist, the overall cost–benefit trade-off may be unfavorable in practical deployment.

## C Ethics Statement

This work is a technical diagnostic study on existing open-source T2V diffusion models. It does not involve human subjects, private data collection, or crowdsourced annotation. We acknowledge the potential societal impact of generative video models (e.g., misinformation or bias), which stems from the pre-trained backbones used in our analysis rather than the proposed initialization method itself.

## D Reproducibility Statement

To ensure reproducibility, we build our experiments upon the open-source VideoCrafter backbone and the VBench evaluation suite. We have detailed the experimental setup, including the number

Table 4: Prompt-level paired significance analysis over 100 prompts (seed-averaged). Mean $\Delta$ is NPNet minus Baseline.

| Metric | Baseline | NPNet | Mean $\Delta$ | 95% CI (bootstrap) | $p$ (perm.) |
|---|---|---|---|---|---|
| temporal_style | 0.076961 | 0.078716 | +0.001754 | $[-0.000658, 0.004166]$ | 0.1687 |

Table 5: Metric card for the qualitative noise-space analysis.

| Metric | Definition / notes |
|---|---|
| RelDisp | $\|d\|_2/\|z\|_2$ on flattened tensors. |
| CosSim | $\cos(z, z_g)$ on flattened tensors. |
| DirStab | Mean pairwise cosine similarity of unit displacements across $S = 5$ seeds (Eq. 4); computed per prompt, then averaged. |
| $\text{CV}_{\|d\|}$ | $\text{Std}(\|d_s\|_2)/\text{Mean}(\|d_s\|_2)$ over seeds; computed per prompt. |
| EVR1 | Top explained-variance ratio from PCA over $\{d_s\}_{s=1}^{5}$ (Eq. 5); computed per prompt. |
| sp_hf($x$) | Spatial HF power ratio from rFFT2 over $(H, W)$ averaged across $(c, t)$ (Eq. 8). |
| t_hf($x$) | Temporal HF power ratio from rFFT over $T$ averaged across $(c, h, w)$, excluding DC (Eq. 10). |
| tDiffRel($x$) | Relative temporal-difference RMS: $\text{RMS}(\Delta_t x)/\text{RMS}(x)$ with $\Delta_t x(:, t) = x(:, t+1) - x(:, t)$. |

of prompts, random seeds, and evaluation protocols, in Section 4 and the Appendix. The complete code, configuration files, and the list of prompts used in this study will be made publicly available on GitHub upon acceptance.

## E Paired significance tests

To rigorously assess the reliability of the observed improvements, we perform a paired statistical analysis with the text prompt as the independent unit ($N = 100$). Table 4 details the results of both bootstrap confidence interval estimation (10,000 resamples) and a paired sign-flip permutation test. The analysis confirms that while the mean difference is positive, the improvement is not statistically significant at the $\alpha = 0.05$ level ($p \approx 0.17$), indicating that the effect size is small relative to the inter-prompt variance.

## F Additional VBench dimension scores

Table 1 reports global mean VBench scores over 100 prompts for completeness. Statistical claims in the main text rely on prompt-level paired testing (Appendix E, Table 4).

## G Definitions of qualitative noise-space metrics

**Notation and scope.** We analyze standard Gaussian initialization $z$ (`z_T`) and the corresponding golden initialization $z_g$ (`z_T_target`) at the same diffusion timestep. We define the displacement

$$dz_g - z. \tag{3}$$

Unless stated otherwise, statistics are aggregated over 100 prompts $\times$ 5 seeds. All frequency ratios use a normalized threshold $\rho = 0.25$; temporal ratios exclude the DC bin.

**Per-prompt aggregation over seeds.** DirStab/CV/EVR1 are computed *per prompt* over the 5 seeds, and then aggregated by mean/median across prompts.

### G.1 Directional stability and low-rank structure

For a prompt, let $\{d_s\}_{s=1}^{S}$ be displacements across $S = 5$ seeds, and define unit directions $u_s = d_s/\|d_s\|_2$. Directional Stability is

$$\text{DirStab} \frac{2}{S(S-1)} \sum_{1 \le i < j \le S} \langle u_i, u_j \rangle. \tag{4}$$

To quantify low-rank structure, we run PCA on the set of flattened $\{d_s\}$ (after centering across seeds). If $\{\lambda_i\}$ are PCA eigenvalues, the explained variance ratio of the top component is

$$\mathrm{EVR1}\frac{\max_i \lambda_i}{\sum_i \lambda_i}. \tag{5}$$

### G.2 Spatial high-frequency ratio

For $x \in \mathbb{R}^{C\times T\times H\times W}$, we compute a 2D real FFT over $(H, W)$ for each channel and time:

$$X(c,t,\omega_h,\omega_w) = \mathrm{rFFT2}\big(x(c,t,:,:)\big), \quad P = |X|^2. \tag{6}$$

We average power over $(c, t)$ to obtain $\bar{P}(\omega_h, \omega_w)$ and define a normalized radial spatial frequency

$$r(\omega_h,\omega_w) = \sqrt{\left(\frac{\min(\omega_h, H-\omega_h)}{H/2}\right)^2 + \left(\frac{\omega_w}{W/2}\right)^2}. \tag{7}$$

With threshold $\rho = 0.25$, the spatial HF ratio is

$$\mathrm{sp\_hf}(x)\frac{\sum_{r\geq\rho}\bar{P}}{\sum\bar{P}}. \tag{8}$$

### G.3 Temporal high-frequency ratio and temporal-difference metrics

For temporal HF ratio, we apply a 1D real FFT along time $T$ for each $(c, h, w)$:

$$Y(c,k,h,w) = \mathrm{rFFT}\big(x(c,:,h,w)\big), \quad Q(k) = \mathbb{E}_{c,h,w}\big[|Y(c,k,h,w)|^2\big], \tag{9}$$

where $k = 0, \ldots, K-1$ and $K = T/2 + 1$. Excluding the DC bin $k = 0$, define normalized frequency $f_k = k/(K-1)$ and threshold $\rho_t = 0.25$:

$$\mathrm{t\_hf}(x)\frac{\sum_{k>0,\, f_k\geq\rho_t} Q(k)}{\sum_{k>0} Q(k)}. \tag{10}$$

To measure temporal discontinuity, we compute first-order differences $\Delta_t x(:,t,:,:) = x(:,t+1,:,:) - x(:,t,:,:)$. We report

$$\mathrm{tDiffRMS}(x)\sqrt{\mathbb{E}\big[(\Delta_t x)^2\big]}, \qquad \mathrm{tDiffRel}(x)\frac{\mathrm{tDiffRMS}(x)}{\sqrt{\mathbb{E}[x^2]}}. \tag{11}$$