FRED: FREQUENCY-RESOLVED DISTILLATION FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step distillation accelerates video generation by reducing the number of denoising evaluations needed at inference. Distribution-matching approaches primarily achieve this by aligning teacher and student distributions in latent space, but such alignment does not explicitly assign teacher guidance to individual spatial frequency bands. To investigate this issue, we analyze generated and real videos in the spatial frequency domain and observe distinct discrepancies in their low- and high-frequency components. These discrepancies motivate *frequency-resolved distillation* with dedicated supervision for low- and high-frequency predictions. Specifically, we decompose the outputs of a frozen teacher model and a trainable fake score model into two frequency bands while retaining the complete input context, and the difference within each band supervises the corresponding component of the same native student prediction, extracted using a non-learnable filter. We further introduce overlap-aware loss weighting to regulate supervision in the transition region where the low- and high-frequency components overlap. Our method adds band-residual losses to the training code and incurs little training overhead, while preserving the original sampler and leaving the number of denoising steps and inference cost unchanged. Experiments across text-to-video, image-to-video, and autoregressive video generation tasks demonstrate improved generation quality and closer spectral agreement with real videos.

1 INTRODUCTION

Diffusion-based video generation has attracted growing interest for content creation, simulation, and prediction (Wan Team et al., 2025; Liao et al., 2026; Hu et al., 2023). However, current video generation remains computationally expensive since it requires repeated denoising of high-dimensional video representations. This iterative denoising process results in substantial sampling latency, limiting applications that require fast responses. Distillation provides a promising solution by transferring the generative capability of a teacher model to a student model that requires fewer sampling steps, thereby reducing inference cost while preserving the quality of the original model (Salimans & Ho, 2022; Song et al., 2023). Distribution-matching methods achieve this by aligning the student's distribution with the target distribution in latent space (Yin et al., 2024b;a; Chen et al., 2026).

To examine the spatial statistics of distilled outputs, we compare the spatial spectra of student-generated and real videos. Low frequencies capture slowly varying patterns over broad spatial regions, whereas high frequencies characterize rapid local variations, providing complementary views of information at different spatial scales. Our analysis reveals a clear mismatch in how spatial variation is distributed across these frequency scales. Although low- and high-frequency components do not strictly correspond to structure and texture, respectively, the mismatch suggests that information at different spatial scales may be unevenly preserved, potentially degrading visual quality and consistency in distilled outputs.

This spectral analysis motivates us to explicitly address discrepancies at different spatial scales during distillation, rather than leaving their supervision entangled within a shared training signal. However, simply decomposing the prediction into frequency components does not separate their supervision when the loss is applied only to the recombined prediction. This motivates band-specific supervision, where the corresponding components of the same native student prediction receive dedicated guidance while the student retains the complete input context.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce *frequency-resolved distillation* to associate each frequency band with dedicated distribution-matching guidance through non-learnable complementary Gaussian operators applied to a single native student output. We supervise each band-specific prediction using the difference between frequency-filtered predictions from a frozen teacher model and a trainable fake score model, allowing low- and high-frequency predictions to receive distinct supervision. We further introduce overlap-aware loss weighting to balance supervision in the overlapping region between low- and high-frequency components. Notably, our method requires only minimal modifications to the existing distillation pipeline, adding band-residual losses to the state-of-the-art DFD framework (Chen et al., 2026) and incurring negligible additional training overhead. We preserve the original sampler and number of denoising steps, with unchanged inference cost.

Our contributions are:

- We use spatial spectral probes to characterize the mismatch between real and student-generated videos, motivating explicit supervision at different spatial scales.
- We introduce *frequency-resolved distillation*, which explicitly decomposes distribution-matching guidance across frequency bands on one native student output and employs overlap-aware loss weighting to coordinate learning in the overlapping frequency region.
- We evaluate our method across text-to-video, image-to-video, and autoregressive video generation tasks. The results demonstrate consistent improvements over the evaluated baselines in visual quality, visual consistency, and temporal coherence.

2 Related Work

Video generation tasks. Video generation research spans text-to-video (T2V) and image-to-video (I2V) synthesis, alongside autoregressive generation. Latent video diffusion (Blattmann et al., 2023b) underlies T2V systems such as ModelScopeT2V (Wang et al., 2023a), CogVideoX (Yang et al., 2025), OpenSora (Zheng et al., 2024), and Wan (Wan Team et al., 2025), which we use for text- and image-conditioned generation. I2V conditions on a reference frame and preserves subject identity and background while adding motion; Stable Video Diffusion (Blattmann et al., 2023a) and Wan's I2V variants are representative. Autoregressive models generate block by block, so the training history is part of the task: Diffusion Forcing uses independent per-token noise (Chen et al., 2024), CausVid distills a bidirectional teacher into a causal student (Yin et al., 2025), Self Forcing trains on self-generated histories (Huang et al., 2025b), One-Forcing adds one-step adversarial supervision (Feng et al., 2026), and Causal-rCM combines teacher-forcing consistency with self-forcing distribution matching (Zheng et al., 2026a). Mode Seeking meets Mean Seeking combines flow matching with distribution-matching distillation for long-range coherence and local realism (Cai et al., 2026). Our method keeps the full text or image context and assigns distribution-matching guidance to separate low- and high-frequency components of the prediction.

Distillation for video generation. Distillation reduces the sampling cost of diffusion models by transferring their generative capability to few-step students. Video distillation uses consistency training (Wang et al., 2023b; Li et al., 2024) or adversarial objectives (Lin & Yang, 2024; Zhang et al., 2024; 2025). DMD aligns the student distribution with a pretrained teacher, while DMD2 improves training stability and sample quality (Yin et al., 2024b;a). DP-DMD uses first-step teacher prediction targets to preserve diversity (Wu et al., 2026). Data-Forcing Distillation introduces real-data teacher queries to improve fidelity and diversity (Chen et al., 2026). rCM combines consistency and score distillation (Zheng et al., 2026b), while TMD uses conditional flow heads (Nie et al., 2026). Adaptive Video Distillation introduces spatially adaptive regression and temporal regularization (You et al., 2026). Building on distribution-matching distillation, we study how to assign teacher guidance to different spatial frequency components of the student prediction.

Frequency-aware generation and distillation. Frequency-aware generation uses spectral losses (Jiang et al., 2021), adaptation (Yin et al., 2026), noise initialization (Wu et al., 2024; Yuan et al., 2025), long-video attention (Lu et al., 2024), and classifier-free guidance caching (Lyu et al., 2025). Concurrent work FreqForcing stabilizes autoregressive rollouts through spectral self-anchoring at inference (Li et al., 2026). Within distillation, FRAMER transfers frequency-separated

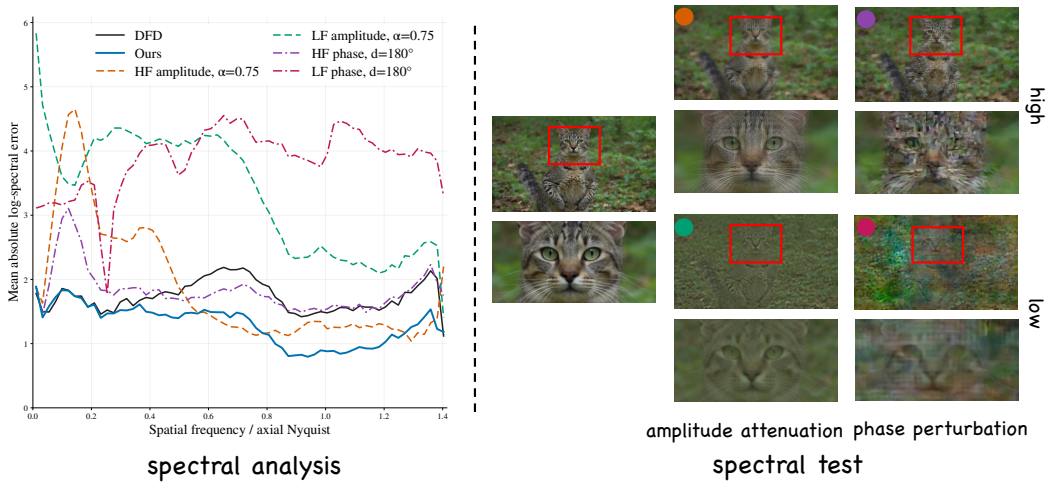


Figure 1: Spatial-frequency diagnostics. Left: radial power-spectrum errors relative to reference frames. Right: decoded results before and after high- and low-frequency amplitude and phase interventions on generated latents, with matched crops.

features between network layers for image super-resolution (Choi et al., 2026), GPD applies a time-gated spatiotemporal high-frequency loss to teacher-refined targets (Liang et al., 2026), and FreqDistill studies supervision over three disjoint spatiotemporal bands in controlled simulations (Sharkey, 2026). FReD decomposes the predictions of a frozen teacher model and a trainable fake score model into soft spatial bands, uses their differences to supervise band-specific components of a single native student output, and weights supervision in the overlapping region between the bands.

3 Frequency-Domain Analysis of Distillation

Full-prediction supervision combines errors across spatial scales, leaving their individual contributions implicit. To motivate frequency-specific supervision, we examine spectral deviations from reference videos and how changing latent amplitude or phase in different bands affects decoded appearance.

Spatial spectral comparison. Figure 1 (left) measures radial power-spectrum errors between generated frames and their corresponding reference frames. We convert frames to luminance, subtract the spatial mean, apply a two-dimensional Hann window, and aggregate power into 64 radial bins. For each evaluated case, we compute the absolute $\log_2$ power ratio within each bin and then average the errors across cases. Our method reduces the mean radial power-spectrum error from 1.7295 to 1.2988 compared with DFD, indicating closer agreement with the reference spectra. The protocol and scope of this comparison are detailed in Appendix C.1.

Controlled latent interventions. Figure 1 (right) applies high- and low-frequency amplitude attenuation and phase perturbations directly to the same native generated video latent from Ours. The unperturbed decoding provides the visual baseline. Non-learnable complementary Gaussian masks ($\sigma_f = 0.35$; Section 4.1) modulate each intervention in the spatial frequency domain. We use amplitude attenuation of $\alpha = 0.75$ and a nominal phase bound of $180°$, preserving the latent DC component. Phase perturbations preserve latent spectral amplitudes and energy, but can change decoded-image power spectra through the nonlinear VAE decoder, as also shown in the left panel. Shared crops reveal local changes in the cat example. Intervention details are provided in Appendix C.2.

Implications for distillation. Amplitude interventions illustrate band-dependent changes in fine detail and broader appearance. Phase interventions alter decoded appearance despite preserving latent spectral amplitudes and energy, showing that amplitude preservation alone is insufficient to

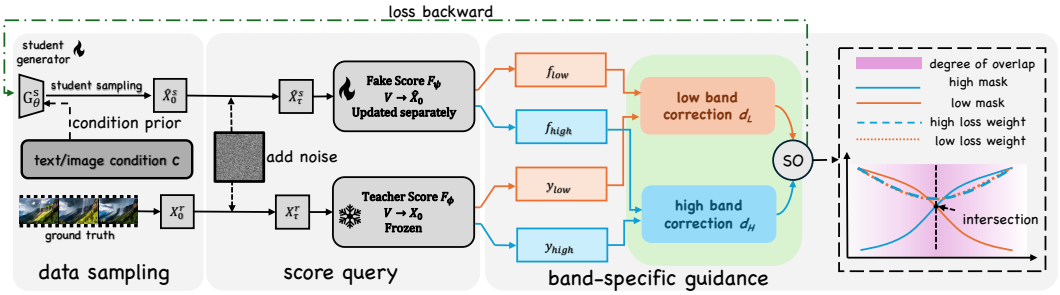


Figure 2: Overview of frequency-resolved distillation. Band-specific guidance supervises band-specific components of the student's complete prediction, while soft-overlap weighting adjusts their contributions in the shared transition region. Both loss paths update the same student output head and backbone.

preserve visual structure. These observations motivate supervision of complete band-specific predictions, whose residuals are sensitive to both amplitude and phase relationships. The contributions of band-specific supervision and overlap weighting are evaluated through method ablations.

4 Frequency-Resolved Distillation

The preceding analysis motivates supervising complete band-specific predictions to capture both amplitude and phase relationships. We introduce band-specific guidance for low and high frequencies, followed by a shared weighting rule that reduces their overlapping contributions. Figure 2 summarizes the overall pipeline.

4.1 Band-Specific Guidance

We build on the DFD student-update procedure (Chen et al., 2026) and replace its full-prediction loss with frequency-specific supervision. The student's single native output head produces the complete clean-latent prediction $\hat{X}_0^s$ shown in Figure 2; we denote this prediction by $g \equiv \hat{X}_0^s$. All frequency operations act in VAE latent space.

For a latent of shape $B \times C \times T \times H \times W$, define the non-learnable spatial filters:

$$P_b(x) = \mathrm{Re}\,\mathcal{F}_{xy}^{-1}(M_b \odot \mathcal{F}_{xy}(x)), \qquad M_L = e^{-\rho^2/(2\sigma_f^2)}, \quad M_H = 1 - M_L, \tag{1}$$

where $b \in \{L, H\}$ and $\mathcal{F}_{xy}$ is the orthonormal spatial fast Fourier transform (FFT) (Cooley & Tukey, 1965), applied per channel and frame. For f_x, f_y in cycles per latent pixel, the normalized radius satisfies $\rho^2 = \min\{(f_x^2 + f_y^2)/0.5, 1\}$.

We choose the Gaussian width so that the low- and high-frequency masks have equal responses at a prescribed radius $\rho_\star$. For the complementary Gaussian masks in Equation 1, this condition gives:

$$M_L(\rho_\star) = M_H(\rho_\star) = \frac{1}{2} \quad \Longleftrightarrow \quad \rho_\star = \sigma_f\sqrt{2\ln 2}. \tag{2}$$

We place this crossover at $\rho_\star = 0.35\sqrt{2\ln 2} \approx 0.412$, giving $\sigma_f = \rho_\star/\sqrt{2\ln 2} = 0.35$. The band predictions are:

$$g_L = P_L(g), \qquad g_H = P_H(g), \qquad g_L + g_H = g. \tag{3}$$

We retain the model queries and shared per-sample normalization from distribution-matching guidance (Yin et al., 2024b;a; Chen et al., 2026). Let f_{fake} and y denote the fake-score and classifier-free guidance (CFG) (Ho & Salimans, 2022) teacher predictions, and a the shared normalization scale. We form a detached correction in each band from their filtered prediction difference:

$$d_b = a\big(P_b(f_{\text{fake}}) - P_b(y)\big), \qquad d = d_L + d_H. \tag{4}$$

The sum d is the normalized, stop-gradient complete guidance correction. By linearity, $d = a(f_{\text{fake}} - y)$ and $d_b = P_b(d)$.

Algorithm 1 Replacing DFD's two-line student loss (Chen et al., 2026) with five lines of band-specific supervision. The full student update is provided in Algorithm 2.

Shared inputs: full student prediction g and detached DFD guidance d from Equation 4.

DFD: full-prediction loss (2 lines)

```
r = g - (g - d).detach()
loss = 0.5 * mean(r**2)
```

FReD: band-specific loss (5 lines)

```
fy, fx = freq_grid(g)
low = exp(-(fy**2 + fx**2) / 0.35**2)
bands = lambda x: stack([IFFT(m*FFT(x)).real for m in (low,1-low)])
r = bands(g) - (bands(g) - bands(d)).detach()
loss = 0.5 * sum(mean(rb**2) for rb in r)
```

`FFT`/`IFFT` are orthonormal spatial transforms; `stack` adds a leading band dimension. The model queries and shared normalization are unchanged.

For each band, we define a detached pseudo-target $\widetilde{g}_b$ and the prediction residual r_b:

$$\widetilde{g}_b = \mathrm{sg}(g_b - d_b), \qquad r_b = g_b - \widetilde{g}_b, \quad b \in \{L, H\}. \tag{5}$$

Here sg denotes stop-gradient; gradients pass only through the student prediction $g_b = P_b(g)$.

The band-specific student objective is:

$$\mathcal{L}_{\text{band}} = \frac{1}{2} \sum_{b \in \{L,H\}} \text{mean}[r_b^2], \tag{6}$$

where the mean covers all batch and latent dimensions. Both losses update the same output head and backbone through P_b. With $N = BCTHW$, their combined gradient is:

$$\frac{\partial \mathcal{L}_{\text{band}}}{\partial g} = \frac{1}{N} \sum_b P_b^* r_b \approx \frac{1}{N} \mathcal{F}_{xy}^{-1} \left[\left(M_L^2 + M_H^2 \right) \odot \mathcal{F}_{xy}(d) \right]. \tag{7}$$

The real symmetric masks give $P_b^* = P_b$. The squared masks arise from their action on both the guidance and the returning gradient, showing how band-specific supervision reshapes the frequency distribution of the update. At inference, the native complete prediction g directly drives the original DFD sampler. Algorithm 1 summarizes the differences between our student update and DFD.

4.2 Soft-Overlap Loss Weighting

The Gaussian decomposition creates a smooth transition region in which spatial frequencies contribute to both band losses. To retain this smooth decomposition while placing greater relative emphasis on frequencies predominantly assigned to either band, we introduce Soft-Overlap Loss Weighting (SO). SO progressively attenuates the shared transition region according to mask overlap, using the same weighting rule for both bands.

For each spatial frequency ω, we quantify mask overlap as:

$$O_\omega = 4 M_L(\omega) M_H(\omega), \qquad O_\omega \in [0, 1]. \tag{8}$$

The overlap reaches one when $M_L = M_H = 0.5$ and approaches zero when either mask approaches zero. To attenuate frequencies in proportion to this overlap, we use the shared weight:

$$W_L(\omega) = W_H(\omega) = W(\omega) = 2^{-\gamma O_\omega}, \qquad \gamma = 1. \tag{9}$$

Thus both bands receive weight $1/2$ at maximum overlap, while the weight approaches one as overlap vanishes. The parameter γ controls the common attenuation strength, and the weighting is invariant to exchanging the low- and high-frequency masks.

Using the band residuals r_b from Section 4.1, we define:

$$\mathcal{L}_{\text{SO}} = \frac{1}{2} \sum_{b \in \{L,H\}} \text{mean}_{B,C,T,\omega} \left[W(\omega) \left| \mathcal{F}_{xy}(r_b)(\omega) \right|^2 \right]. \tag{10}$$

Here $\mathcal{F}_{xy}$ uses orthonormal normalization. SO retains the band-specific guidance and detached pseudo-targets, and weights the residual energy at each spatial frequency. Setting $\gamma = 0$ recovers the band loss in Equation (6) by Parseval's identity.

To characterize its effect on the student update, define:

$$K_{\text{base}} = M_L^2 + M_H^2, \qquad K_{\text{SO}} = W(M_L^2 + M_H^2) = W K_{\text{base}}. \tag{11}$$

For the normalized, detached guidance d defined in Section 4.1 and $N = BCTHW$, the gradient at the complete prediction is:

$$\frac{\partial \mathcal{L}_{\text{SO}}}{\partial g} = \frac{1}{N} \mathcal{F}_{xy}^{-1}[K_{\text{SO}} \odot \mathcal{F}_{xy}(d)]\,. \tag{12}$$

The squared masks arise from filtering the guidance and propagating the gradient through the student-side filters. The shared weight applies the same attenuation to both loss paths at each spatial frequency, preserving their relative contributions under the base band loss. At the mask crossover, $K_{\text{base}} = 1/2$ and $K_{\text{SO}} = 2^{-(\gamma+1)}$, which equals $1/4$ for our choice $\gamma = 1$. Where overlap vanishes, K_{SO} approaches K_{base}. Both loss paths update the same student output head and backbone. Implementation details are provided in Appendix B.1.

5 Experiments

We compare our method with DMD2, DP-DMD, and DFD on text-to-video, autoregressive, and image-to-video generation using six VBench metrics for quality, consistency, and motion (Huang et al., 2024). Runtime is measured on NVIDIA H800 GPUs with 80 GB of memory. Training and evaluation details are provided in Appendix B.2.

5.1 Text-to-Video Generation

Experimental Settings. We use Wan2.1-T2V-1.3B (Wan Team et al., 2025), assessing video quality, CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) feature diversity (video-mean and per-frame), and camera diversity via endpoint and trajectory distances.

Experimental Results. Figure 3 shows clearer appearance and finer details than DFD; additional examples are in Appendix E.1. Table 1 shows that our method achieves the highest average VBench score, with strong gains in imaging and aesthetic quality. It also improves all four visual-diversity metrics and both camera-diversity metrics, indicating greater variation alongside improved visual quality. A matched-budget comparison shows that further training DFD leaves its overall quality largely unchanged, while our method improves all six reported quality metrics (Appendix B.2, Table 5).

5.2 Autoregressive Video Generation

Experimental Settings. We use the causal variant of Wan 1.3B (Wan Team et al., 2025) for autoregressive video generation, conditioning each block on preceding content through a KV cache. We evaluate video quality and training and inference times using the shared protocol.

Experimental Results. Our method improves all six VBench metrics over the baselines (Table 2). Figure 4 shows clearer appearance and richer details, while the temporal scores indicate improved consistency and motion smoothness during autoregressive generation.

5.3 Image-to-Video Generation

Experimental Settings. We use Wan2.1-Fun-1.3B-InP-Diffusers (Alibaba PAI, 2025), conditioning generation on an input image. Alongside the six VBench metrics, we evaluate first-frame subject and background consistency to measure preservation of the input scene.

Experimental Results. Our method improves both video quality and first-frame consistency over the baselines (Table 3), balancing generated detail with fidelity to the input image. Figure 7 (Appendix E.2) illustrates clearer subject details while preserving the input scene's appearance. Across all three tasks, training and inference times remain comparable to those of DFD.

Table 1: Text-to-video evaluation across video quality, video diversity, and camera pose diversity. For quality and diversity metrics, higher is better; bold indicates the highest score. Lower runtime is better.

	Video Quality							Runtime (s)	
Method	Subj. Consist.	Bkgd. Consist.	Temp. Flicker.	Image Quality	Aesth. Quality	Motion Smooth.	Avg.	Train	Infer.
DMD2	95.62	95.03	97.87	67.98	61.29	98.02	85.97	43.656	11.554
DP-DMD	95.23	94.90	97.90	68.01	62.31	98.70	86.18	78.408	11.533
DFD	95.41	95.14	97.88	68.36	63.12	98.84	86.46	30.630	11.672
Ours	**95.92**	**95.53**	**98.03**	**70.78**	**64.01**	**99.08**	**87.23**	30.530	11.662

	Video Diversity				Camera Pose Diversity	
Method	CLIP (Mean)	CLIP (Per-frame)	DINO (Mean)	DINO (Per-frame)	Endpoint Distance	Trajectory Distance
DMD2	0.0719	0.0977	0.1534	0.2021	3.8584	1.8973
DP-DMD	0.0759	0.1139	0.1498	0.2019	3.1272	1.6557
DFD	0.0723	0.1027	0.1651	0.2135	3.7108	2.2108
Ours	**0.0773**	**0.1150**	**0.1748**	**0.2194**	**5.3239**	**4.3194**

Table 2: Autoregressive video generation results. For quality metrics, higher is better; bold indicates the highest score. Lower runtime is better.

	Video Quality							Runtime (s)	
Method	Subj. Consist.	Bkgd. Consist.	Temp. Flicker.	Image Quality	Aesth. Quality	Motion Smooth.	Avg.	Train	Infer.
DMD2	90.83	92.73	95.17	65.37	59.95	97.46	83.59	65.955	7.587
DP-DMD	90.36	92.80	94.86	65.17	59.82	97.35	83.39	87.143	7.592
DFD	91.15	93.07	95.77	65.77	60.71	97.78	84.04	65.670	7.613
Ours	**91.55**	**93.29**	**96.34**	**66.51**	**61.27**	**98.08**	**84.51**	65.905	7.631

Table 3: Image-to-video results. FF Consist.: first-frame subject/background consistency (I2V-S/I2V-B). Avg.: six video-quality metrics. Higher scores and lower runtime are better; best scores are bold.

	Video Quality							FF Consist.		Runtime (s)	
Method	SC	BC	TF	IQ	AQ	MS	Avg.	I2V-S	I2V-B	Train	Infer.
DMD2	90.06	92.58	95.24	65.97	61.83	98.37	84.01	94.84	96.76	50.281	15.855
DP-DMD	94.90	95.62	96.83	69.33	65.81	98.55	86.84	98.20	98.69	85.339	15.593
DFD	96.44	96.69	98.45	69.94	66.44	99.04	87.83	98.90	99.22	37.207	15.377
Ours	**97.40**	**97.27**	**98.94**	**70.06**	**66.65**	**99.21**	**88.26**	**99.03**	**99.29**	37.282	15.346

5.4 Component Analysis

Effect of Soft-Overlap Loss Weighting. Table 4 compares training with and without SO. SO generally improves video quality and consistency while maintaining comparable temporal performance, supporting overlap attenuation as a way to balance appearance and motion. Experimental details are in Appendix D.1.

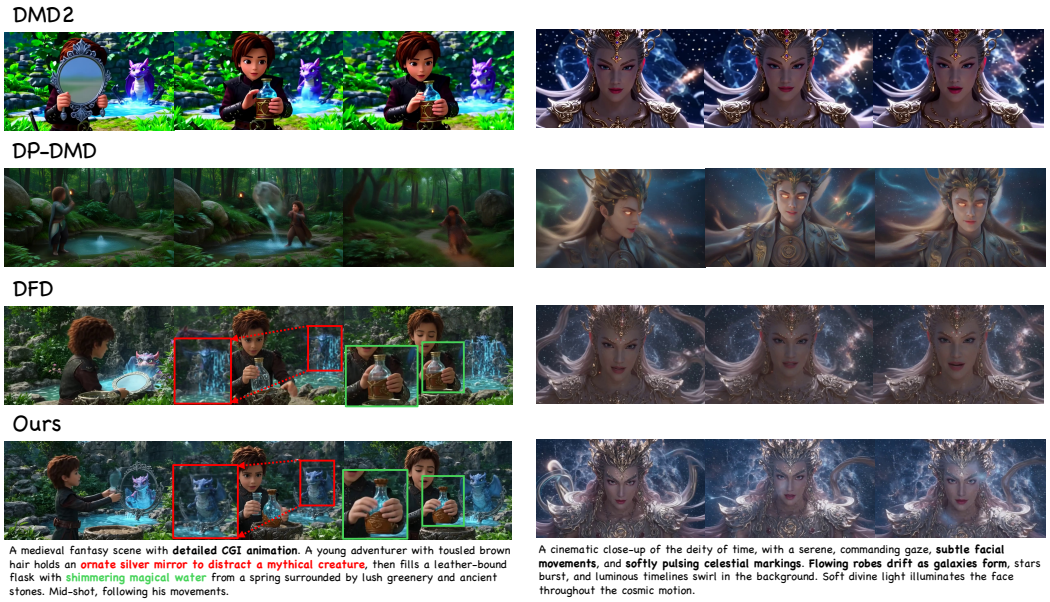


Figure 3: Text-to-video qualitative comparison for a fantasy scene and a close-up of the deity of time. Our method best preserves low-frequency global structure and high-frequency texture details among the compared methods.



Figure 4: Autoregressive video comparisons for a movie-theater scene and furry characters playing a video game. Our method produces more harmonious colors and more natural lighting while better following the prompts.

Effect of Band-Specific Supervision. We compare Low-only, High-only, and Low+High supervision with SO disabled and $\sigma_f = 0.35$. Figure 5 and Table 6 show that removing either band compromises video quality: low-frequency supervision supports the overall scene appearance, while high-frequency supervision helps preserve fine details. Combining both yields more consistent videos with better visual quality, supporting their complementary roles in distillation. Experimental details are in Appendix D.2.

Table 4: Effect of Soft-Overlap Loss Weighting (SO). Steps denote continued training. All metrics are higher-is-better; bold marks the better score within each matched-step pair.

Steps	SO	Subject Consist.	Bkgd. Consist.	Aesthetic Quality	Imaging Quality	Temporal Flicker.	Motion Smooth.
40	×	95.837	95.482	63.907	70.569	**98.031**	**99.101**
40	✓	**95.924**	**95.534**	**64.006**	**70.775**	98.026	99.081
60	×	95.872	95.443	63.819	70.753	**98.001**	**99.014**
60	✓	**95.928**	**95.497**	**63.929**	**70.797**	97.969	98.990
80	×	95.771	95.403	63.480	**69.813**	**97.879**	**98.822**
80	✓	**95.798**	**95.517**	**63.711**	69.803	97.866	98.821
100	×	95.739	95.370	63.264	68.923	98.027	98.899
100	✓	**95.764**	**95.427**	**63.384**	**68.982**	**98.030**	**98.903**

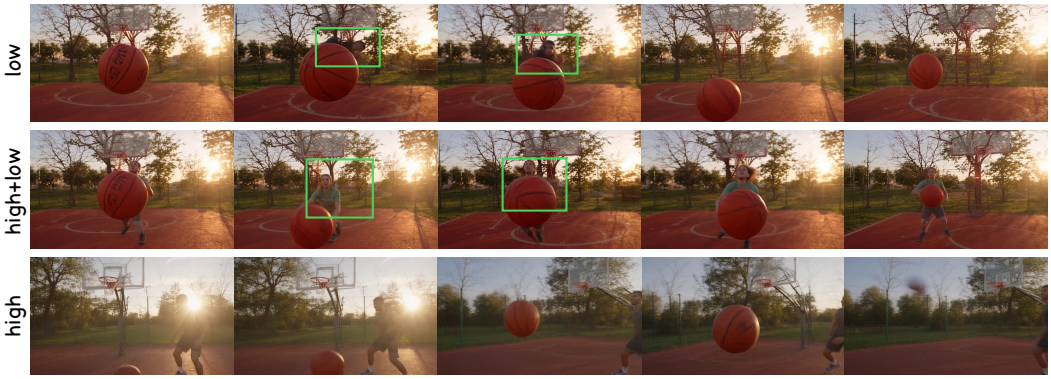


Figure 5: Band-specific supervision ablation. Low+High better preserves low-frequency appearance, including coherent lighting and colors, while maintaining fine details. Low-only exhibits flickering fine details across frames, whereas High-only produces less coherent lighting and poorer color rendition. Prompt: "Slow-motion shot of a basketball player dribbling on an outdoor court at golden hour."

Effect of Frequency Bandwidth. We vary the Gaussian bandwidth $\sigma_f \in \{0.1, 0.35, 0.6\}$ with SO disabled. Different bandwidths yield tradeoffs between visual quality, consistency, and temporal stability, with $\sigma_f = 0.35$ providing a favorable balance. Results and experimental details are in Appendix D.3, Table 7.

6 Limitations and Conclusion

Our method uses non-learnable spatial Gaussian filters and a symmetric overlap-weighting rule. Spatial frequency does not uniquely separate semantic structure from texture, and the complementary bands share a transition region. These spatial losses do not explicitly model temporal frequencies. In image-to-video generation, output quality may be limited by the quality of the conditioning first frame. Moreover, our diagnostics and ablations suggest that low- and high-frequency components affect video quality differently, but we have not yet explored distinct methods tailored to the specific roles of each frequency band.

Conclusion. We present frequency-resolved distillation for video generation, combining band-specific supervision of a single student output with soft-overlap loss weighting. Experiments on text-to-video, autoregressive, and image-to-video generation demonstrate improved quality with little training overhead and unchanged sampling procedures and inference cost.

AI USE STATEMENT

An LLM-based assistant was used to assist with literature searches and to draft and polish the English text. The assistant did not perform model training, model inference, or additional experiments. All numerical results were provided by the authors, who take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

Video generation models may be misused to create deceptive media or impersonate individuals. Any release of models or generated outputs should therefore consider content provenance, clearly identify synthetic content, and respect applicable restrictions on source materials. This study does not involve human-subject evaluation and does not reproduce private videos. Any future qualitative or preference studies involving participants should obtain informed consent, protect the privacy of identifiable individuals, and ensure fair participant treatment.

REPRODUCIBILITY STATEMENT

The frequency decomposition, student architecture, and distillation objectives are specified in Section 4, with the student update summarized in Algorithm 1. Appendix A provides the mathematical derivations. Appendix B describes the implementation, training-data sources and held-out splits, model initialization, optimizer hyperparameters, update schedules, and runtime measurement hardware. It also documents the main evaluation prompt sources, numbers of generation seeds, quality and diversity metrics, and score aggregation. Appendix C details the spatial-frequency analysis and latent-intervention protocols. Appendix D specifies the ablation configurations, training budgets, and inference settings, while Appendix E provides additional qualitative comparisons.

REFERENCES

Alibaba PAI. *Wan2.1-Fun-1.3B-InP*, 2025. URL https://huggingface.co/alibaba-pai/Wan2.1-Fun-1.3B-InP. Hugging Face model release.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align Your Latents: High-Resolution Video Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, 2023b.

Shengqu Cai, Weili Nie, Chao Liu, Julius Berner, Lvmin Zhang, Nanye Ma, Hansheng Chen, Maneesh Agrawala, Leonidas Guibas, Gordon Wetzstein, and Arash Vahdat. Mode Seeking meets Mean Seeking for Fast Long Video Generation. In *International Conference on Machine Learning*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *Advances in Neural Information Processing Systems*, pp. 24081–24125, 2024.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-Forcing Distillation: Restoring Diversity and Fidelity in Few-Step Video Generation. *arXiv preprint arXiv:2606.18478*, 2026.

Seungho Choi, Jeahun Sung, and Jihyong Oh. FRAMER: Frequency-Aligned Self-Distillation with Adaptive Modulation Leveraging Diffusion Priors for Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23451–23461, 2026.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-Forcing: Towards Stable One-Step Autoregressive Video Generation. *arXiv preprint arXiv:2605.23458*, 2026.

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

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A Generative World Model for Autonomous Driving. *arXiv preprint arXiv:2309.17080*, 2023.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixé, and Sanja Fidler. ViPE: Video Pose Engine for 3D Geometric Perception. *arXiv preprint arXiv:2508.10934*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, pp. 167283–167308, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal Frequency Loss for Image Reconstruction and Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13919–13929, 2021.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback. In *Advances in Neural Information Processing Systems*, pp. 75692–75726, 2024.

Jiatong Li, Leo Liang, Linghe Kong, and Yulun Zhang. FreqForcing: Autoregressive Long Video Generation via Spectral Self-Anchoring. *arXiv preprint arXiv:2607.27110*, 2026.

Xiao Liang, Yunzhu Zhang, and Linchao Zhu. GPD: Guided Progressive Distillation for Fast and High-Quality Video Generation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1406–1414, 2026.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie Envisioner: A Unified World Foundation Platform for Robotic Manipulation. In *International Conference on Learning Representations*, pp. 88446–88463, 2026.

Shanchuan Lin and Xiao Yang. AnimateDiff-Lightning: Cross-Model Diffusion Distillation. *arXiv preprint arXiv:2403.12706*, 2024.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. FreeLong: Training-Free Long Video Generation with SpectralBlend Temporal Attention. In *Advances in Neural Information Processing Systems*, pp. 131434–131455, 2024.

Zhengyao Lyu, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K. Wong. FasterCache: Training-Free Video Diffusion Model Acceleration with High Quality. In *International Conference on Learning Representations*, pp. 33132–33156, 2025.

Weili Nie, Julius Berner, Nanye Ma, Chao Liu, Saining Xie, and Arash Vahdat. Transition Matching Distillation for Fast Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4645–4655, 2026.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*, 2022.

Susan Sharkey. FreqDistill: Spectral Teacher-Student Distillation for Efficient Long-Video Diffusion Transformers. Author-uploaded technical report, ResearchGate, May 10, 2026, 2026. URL https://doi.org/10.13140/RG.2.2.25180.76160.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. In *International Conference on Machine Learning*, pp. 32211–32252, 2023.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope Text-to-Video Technical Report. *arXiv preprint arXiv:2308.06571*, 2023a.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. VideoLCM: Video Latent Consistency Model. *arXiv preprint arXiv:2312.09109*, 2023b.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-Preserved Distribution Matching Distillation for Fast Visual Synthesis. In *International Conference on Machine Learning*, 2026.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging Initialization Gap in Video Diffusion Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 378–394, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations*, pp. 83048–83077, 2025.

Bo Yin, Xiaobin Hu, Xingyu Zhou, Yu He, Peng-Tao Jiang, Yue Liao, Junwei Zhu, Jiangning Zhang, Ying Tai, and Shuicheng Yan. FeRA: Frequency-Energy Constrained Routing for Effective Diffusion Adaptation Fine-Tuning. In *International Conference on Machine Learning*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. In *Advances in Neural Information Processing Systems*, pp. 47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step Diffusion with Distribution Matching Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2025.

Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive Video Distillation: Mitigating Oversaturation and Temporal Collapse in Few-Step Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43429–43439, 2026.

Yunlong Yuan, Yuanfan Guo, Chunwei Wang, Wei Zhang, Hang Xu, and Li Zhang. FreqPrior: Improving Video Diffusion Models with Frequency Filtering Gaussian Noise. In *International Conference on Learning Representations*, pp. 4661–4682, 2025.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. AccVideo: Accelerating Video Diffusion Model with Synthetic Dataset. *arXiv preprint arXiv:2503.19462*, 2025.

Zhixing Zhang, Yanyu Li, Yushu Wu, Yanwu Xu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Dimitris Metaxas, Sergey Tulyakov, and Jian Ren. SF-V: Single Forward Video Generation Model. In *Advances in Neural Information Processing Systems*, pp. 103599–103618, 2024.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rCM: A Unified Teacher-Forcing and Self-Forcing Open Recipe for Autoregressive Diffusion Distillation in Streaming Video Generation and Interactive World Models. *arXiv preprint arXiv:2606.25473*, 2026a.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large Scale Diffusion Distillation via Score-Regularized Continuous-Time Consistency. In *International Conference on Learning Representations*, pp. 2582–2603, 2026b.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All. *arXiv preprint arXiv:2412.20404*, 2024.

A Theoretical Analysis

We use the notation of Section 4: g is the complete student prediction, y is the CFG teacher prediction, f_{fake} is the fake-score prediction, and d_b is the detached guidance in band b. A video latent has shape $B \times C \times T \times H \times W$; frequency transforms act on H, W within each channel and frame.

A.1 Non-Learnable Frequency Operators

The implementation uses the unshifted coordinates returned by `torch.fft.fftfreq`:

$$\rho^2 = \min\left(\frac{f_x^2 + f_y^2}{0.5}, 1\right), \qquad M_L = \exp\left(-\frac{\rho^2}{2\sigma_f^2}\right), \quad M_H = 1 - M_L. \tag{13}$$

The denominator 0.5 is the sum of squared axial Nyquist frequencies. The mask and FFT use the same unshifted ordering. Unless varied in the bandwidth ablation, $\sigma_f = 0.35$. The equal-response radius is $\rho_\star = \sigma_f\sqrt{2\ln 2}$. In Algorithm 1, `freq_grid` supplies these spatial frequency coordinates; on this grid, $\rho^2/(2\sigma_f^2) = (f_x^2 + f_y^2)/\sigma_f^2$, giving the compact mask expression used there.

Complementarity and guidance recovery. In exact arithmetic, the complementary filters satisfy:

$$P_L(x) + P_H(x) = \mathcal{F}_{xy}^{-1}[(M_L + M_H) \odot \mathcal{F}_{xy}(x)] = x. \tag{14}$$

For real, even masks, $P_b^* = P_b$. Their squared and cross-composed operators have masks M_b^2 and $M_L M_H$, respectively; the filters are generally neither idempotent nor orthogonal. With shared detached normalization a, the band corrections satisfy:

$$d_L + d_H = a(P_L + P_H)(f_{\text{fake}} - y) = a(f_{\text{fake}} - y) = d. \tag{15}$$

Thus their sum preserves the complete guidance signal, while applying separate band losses changes the gradient received by the student.

A.2 Gradient of Band-Specific Supervision

Let $N = BCTHW$, $g_b = P_b(g)$, and $r_b = g_b - \text{sg}(g_b - d_b)$. The pseudo-target is treated as a constant during differentiation. For the unweighted band loss, the output gradient is:

$$\frac{\partial \mathcal{L}_{\text{band}}}{\partial g} = \frac{1}{N}\sum_b P_b^* r_b = \frac{1}{N}\mathcal{F}_{xy}^{-1}[(M_L^2 + M_H^2) \odot \mathcal{F}_{xy}(d)]. \tag{16}$$

The two mask factors arise from filtering the guidance and backpropagating through the student-side filter.

A.3 Effect of Soft-Overlap Weighting

With $A = \mathcal{F}_{xy}^{-1}\,\text{diag}(W)\mathcal{F}_{xy}$, the weighted objective gives:

$$\frac{\partial \mathcal{L}_{\text{SO}}}{\partial g} = \frac{1}{N}\sum_b P_b^* A r_b. \tag{17}$$

The forward identity $r_b = d_b = P_b(d)$ does not permit replacing the residual in the loss with detached $P_b(d)$ before differentiation: the gradient comes from the live student prediction g_b. After differentiating with the pseudo-target treated as a constant, using this identity and $P_b^* = P_b$ gives:

$$\frac{\partial \mathcal{L}_{\text{SO}}}{\partial g} = \frac{1}{N}\mathcal{F}_{xy}^{-1}[W(M_L^2 + M_H^2) \odot \mathcal{F}_{xy}(d)]. \tag{18}$$

Both bands update the same native output head and backbone:

$$\nabla_\theta \mathcal{L} = \left(\frac{\partial g}{\partial \theta}\right)^\top \frac{\partial \mathcal{L}}{\partial g}. \tag{19}$$

Setting $W = 1$ recovers the unweighted band gradient. With $O = 4M_L M_H$, the unweighted coefficient is $M_L^2 + M_H^2 = 1 - O/2$. Symmetric SO multiplies this coefficient by $2^{-\gamma O}$ (with $\gamma = 1$ in our configuration); it further attenuates the shared frequencies rather than inverting the filtering effect.

A.4 Amplitude and Phase Sensitivity

With an orthonormal spatial FFT, Parseval's identity gives:

$$\frac{1}{2}\sum_b \mathrm{mean}(r_b^2) = \frac{1}{2}\sum_b \mathrm{mean}\,|\mathcal{F}_{xy}(r_b)|^2. \tag{20}$$

This is the same residual loss expressed in two domains. The residual includes both amplitude and phase discrepancies. For $S = A_s e^{i\phi_s}$ and $T = A_t e^{i\phi_t}$, the squared residual is:

$$|S - T|^2 = (A_s - A_t)^2 + 2A_s A_t[1 - \cos(\phi_s - \phi_t)]. \tag{21}$$

B Implementation and Experimental Settings

B.1 Implementation Details

Prediction and conditioning. The student retains the parent's native output head and produces the complete clean-latent prediction g. The non-learnable operators produce $g_L = P_L(g)$ and $g_H = P_H(g)$ without additional learnable heads. T2V retains full text conditioning; I2V additionally retains the first-frame image representation and mask. The AR setting retains its causal context and KV cache. Filtering is applied to predicted video latents, not to the conditioning inputs.

The network output is converted to a clean prediction using its scheduler. For the linear flow convention $u = (1 - \sigma_{t_s})x_0 + \sigma_{t_s}\epsilon$ and $v = \epsilon - x_0$, this conversion is $g = u - \sigma_{t_s} v$. Other prediction parameterizations use their corresponding scheduler conversion.

Teacher and fake-score queries. Let $Q_t(x, \epsilon)$ denote forward noising at query time t. Student time t_s and query time t have distinct roles. The fake-score model queries $Q_t(g, \epsilon)$, while the teacher queries $Q_t(q, \epsilon)$, with the teacher query defined as:

$$q = \begin{cases} z_{\mathrm{real}}, & \text{real-data query,} \\ \mathrm{sg}(g), & \text{generated-data query.} \end{cases} \tag{22}$$

Our configuration chooses the real-data query with probability 0.5, using one random choice per student-update call. Both models share the query time and noise; their clean query inputs differ in the real-data case. The teacher is frozen. Teacher and fake-score predictions are computed without gradients during the student update, and the fake-score model is trained separately.

We follow DFD (Chen et al., 2026) for teacher queries and classifier-free guidance, using a guidance scale of 5 in our Wan-based setup. CFG is applied to the complete teacher prediction before frequency decomposition. We then apply the non-learnable operators P_b to the guided teacher prediction y and the fake-score prediction f_{fake}, obtaining the band-specific predictions used to construct our supervision. The band losses and SO weighting act on this supervision.

Shared normalization and pseudo-targets. We inherit the per-sample guidance normalization from the DFD implementation: the scale a is the reciprocal of the mean absolute difference between the complete student and teacher predictions, with a stabilizing constant of 10^{-6} and stop-gradient applied. In our implementation, these predictions are reconstructed as $\bar{g} = g_L + g_H$ and $\bar{y} = P_L(y) + P_H(y)$. Both bands share the resulting scale rather than being normalized separately. The corrections $d_b = a[P_b(f_{\mathrm{fake}}) - P_b(y)]$ and pseudo-targets $\widetilde{g}_b = \mathrm{sg}(g_b - d_b)$ are detached; the residual $r_b = g_b - \widetilde{g}_b$ retains gradients only through g_b. The resulting student gradient is derived in Appendix A.2. The main formulation uses the symmetric weight $W = 2^{-O_\omega}$ on both band losses.

Student update and numerical implementation. The student produces one complete prediction g and obtains $g_b = P_b(g)$ with non-learnable filters. The same SO loss is used for real-data and generated-data teacher queries; disabling SO gives $\mathcal{L}_{\mathrm{band}}$. No additional full-band loss is included.

The shared weight is a predefined $H \times W$ map broadcast across batch, channels, and time. The implementation clamps $4M_L M_H$ to $[0, 1]$ when constructing the overlap map. Each weight multiplies the residual spectral energy once; equivalently, the residual spectrum can be multiplied by $\sqrt{W}$ before squaring. FFTs are orthonormal and operate over the spatial dimensions only.

FFTs and spectral energies are computed in FP32. Band outputs, pseudo-target construction, and residual subtraction retain the native prediction dtype (BF16 in the reference configuration). Casting a residual to FP32 does not undo rounding in the preceding subtraction. The shared normalizer is computed from reconstructed student and teacher bands. This agrees with normalization from complete predictions in exact arithmetic, but need not be bitwise identical. The identities in the main text describe exact arithmetic.

The symmetric formulation sets both overlap exponents to one. The weighting adds no learnable parameters or model queries. It introduces spectral computation during training, while inference continues to use the native complete prediction.

Algorithm 2 Full student update corresponding to Algorithm 1, comparing frequency-specific supervision with DFD (Chen et al., 2026). Blue marks the original DFD loss; rose pink highlights our modifications.

```
# f/fake/T: student/fake-score/CFG teacher.
# g: full x0 prediction; d: detached guidance.
# u,ts: student input/time; t,eps: query time/noise.
# real,c: data/condition; FFT/IFFT: spatial transforms.
def student_update(u, ts, t, eps, real, c):
    g = f(u, ts, c)
    q = real if rand() < 0.5 else g.detach()
    with no_grad():
        f_fake = fake(add_noise(g, eps, t), t, c)
        y = T(add_noise(q, eps, t), t, c)
        a = 1 / (mean_per_sample(abs(g - y)) + 1e-6)
        d = a * (f_fake - y)
    # DFD: full-prediction loss.
    # r = g - (g - d).detach()
    # loss = 0.5 * mean(r**2)
    # Ours: band-residual loss.
    fy, fx = freq_grid(g)
    low = exp(-(fy**2 + fx**2) / 0.35**2)
    bands = lambda x: stack([IFFT(m*FFT(x)).real for m in (low,1-low)])
    r = bands(g) - (bands(g) - bands(d)).detach()
    loss = sum(0.5 * mean(rb**2) for rb in r)
    return loss
```

Unweighted control. An arithmetic-matched unweighted control sets $W = 1$ while retaining the same FP32 FFT, squared-magnitude computation, and mean reduction. By Parseval's identity, this recovers the spatial residual MSE in exact arithmetic. Training and evaluation protocols for the three main-text ablations are given in Appendix D.

B.2 Training and Evaluation Settings for T2V, AR, and I2V

We describe the training and evaluation settings for each task below. Our method uses the band-specific supervision and symmetric SO loss in Section 4, with $\sigma_f = 0.35$ and $W = 2^{-O_\omega}$. The teacher remains frozen, and teacher queries use real data with probability 0.5. CFG and shared guidance normalization follow Appendix B.1.

Text-to-video generation. *Training.* We use Wan2.1-T2V-1.3B and initialize the student from the checkpoint released by the DFD authors (Chen et al., 2026). We continue training for 40 iterations on DFD's ViPE-Long data, reserving an additional held-out test split that is excluded from training. At each student update, we choose between a DMD query using the student's generated prediction and a DFD query using real data for the teacher, each with probability 0.5. The fake-score model always queries the generated prediction. Both the student and fake-score model use AdamW with learning rate 10^{-5}, betas $(0.9, 0.999)$, and weight decay 0.01. The global batch size is 16, with one student update for every four fake-score updates; EMA is disabled. The backbone receives the

complete text condition, while frequency decomposition applies to predicted video latents. For the baselines, DMD2 and DFD use checkpoints released by the DFD authors. DP-DMD is initialized from the CityData checkpoint in that release and further fine-tuned on ViPE-Long for 500 iterations.

Evaluation. We use 71 prompts released by the DFD authors (Chen et al., 2026) and 71 prompts released by the One-Forcing authors (Feng et al., 2026). Eight generation seeds per prompt yield 1,136 videos per method. We compare with DMD2, DP-DMD, and DFD using six VBench (Huang et al., 2024) metrics: subject consistency, background consistency, temporal flickering, motion smoothness, aesthetic quality, and imaging quality. Scores are shown on a 0–100 scale, with higher values indicating better performance. Avg. is the arithmetic mean of these six metrics, rather than the official VBench aggregate. Temporal flickering and motion smoothness assess temporal stability rather than motion magnitude. We additionally report diversity based on CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) at the video-mean and per-frame levels, and camera endpoint and trajectory distances. These diversity measures are reported separately from quality. The two prompt sets have equal sample counts, so sample-count-weighted aggregation is equivalent to their arithmetic mean.

Visual diversity. We follow DFD's visual-diversity protocol (Chen et al., 2026), adapted from DP-DMD (Wu et al., 2026). For each of J prompts, compare all $\binom{L}{2} = 28$ pairs of videos generated with $L = 8$ seeds. Let $h^E_{j,i,k}$ be the feature of sampled frame k from video i for prompt j, using encoder E: CLIP-ViT-Large (Radford et al., 2021) or DINOv2-ViT-Large (Oquab et al., 2023). Following DFD, sample $K = 16$ evenly spaced frames and define:

$$\bar{h}^E_{j,i} = \frac{1}{K}\sum_{k=1}^{K} h^E_{j,i,k}, \qquad \delta(u,v) = 1 - \frac{u^\top v}{\|u\|_2\|v\|_2}. \tag{23}$$

The Mean entries in Table 1 use:

$$D^E_{\text{mean}} = \frac{2}{JL(L-1)}\sum_{j=1}^{J}\sum_{1\le i<i'\le L}\delta(\bar{h}^E_{j,i}, \bar{h}^E_{j,i'}). \tag{24}$$

For the Per-frame entries, compare corresponding sampled frames before averaging:

$$D^E_{\text{frame}} = \frac{2}{JL(L-1)K}\sum_{j=1}^{J}\sum_{1\le i<i'\le L}\sum_{k=1}^{K}\delta(h^E_{j,i,k}, h^E_{j,i',k}). \tag{25}$$

Thus Mean pools features before measuring distance, whereas Per-frame averages frame-level distances. Each pair is compared within the same prompt; prompt scores are averaged equally. Higher values indicate greater feature diversity.

Camera-pose diversity. Following DFD (Chen et al., 2026), ViPE (Huang et al., 2025a) estimates camera-to-world poses. Let $p_{j,i,t} \in \mathbb{R}^3$ denote their translations over T_v frames. The reported endpoint and trajectory distances are:

$$D_{\text{end}} = \frac{2}{JL(L-1)}\sum_{j=1}^{J}\sum_{i<i'}\|p_{j,i,T_v} - p_{j,i',T_v}\|_2, \tag{26}$$

$$D_{\text{traj}} = \frac{2}{JL(L-1)T_v}\sum_{j=1}^{J}\sum_{i<i'}\sum_{t=1}^{T_v}\|p_{j,i,t} - p_{j,i',t}\|_2. \tag{27}$$

These compare terminal positions and full translation trajectories across seeds, respectively, then average over prompts.

Matched-budget comparison. We additionally continue training the released T2V DFD checkpoint for 40 iterations using the DFD objective. DFD +40 and our method start from the same released checkpoint and use the same additional training budget. Table 5 compares their VBench results with the released DFD baseline. Continued DFD training leaves the overall quality largely unchanged, whereas our method improves all six metrics.

Runtime is measured on NVIDIA H800 GPUs with 80 GB of memory and reported in seconds. Training time covers one student update and four fake-score updates; inference time covers generation of one complete video. The DMD2 and DP-DMD timing configurations include GAN training,

Table 5: T2V comparison with a matched continued-training budget. DFD +40 and Ours each continue training from the DFD release checkpoint for 40 iterations. All metrics are higher-is-better; Avg. is the mean of the six reported quality metrics.

Method	SC	BC	TF	IQ	AQ	MS	Avg.
DFD release	95.41	95.14	97.88	68.36	63.12	98.84	86.46
DFD +40	95.39	95.07	97.91	68.39	63.14	98.73	86.44
Ours	**95.92**	**95.53**	**98.03**	**70.78**	**64.01**	**99.08**	**87.23**

so auxiliary-update timing covers both fake-score and discriminator computation. DMD2 takes 12.105 seconds for the student update and 31.550 seconds for the four auxiliary updates, with a reported total of 43.656 seconds; DP-DMD takes 35.296 and 43.112 seconds, totaling 78.408 seconds. Their inference times are 11.554 and 11.533 seconds, respectively.

Autoregressive generation. *Training.* We use the causal variant of Wan 1.3B, conditioning each generated block on preceding content through its KV cache. We initialize our student from the DFD release checkpoint and continue training for 70 iterations on DFD's ViPE-Long data, with an additional held-out test split excluded from training. At each student update, the teacher uses a DMD generated-data query or a DFD real-data query with equal probability 0.5; the fake-score model always queries generated data. The student and fake-score optimizers are AdamW with learning rates 2×10^{-6} and 4×10^{-7}, respectively; both use betas $(0, 0.999)$ and weight decay 0.01. The global batch size is 64, the student-to-fake-score update ratio is $1 : 4$, and the EMA decay is 0.99. DMD2 and DFD baselines use checkpoints from the DFD release, while DP-DMD is initialized from the DMD2 checkpoint and trained for 500 additional iterations.

Evaluation. We use 128 prompts released by the One-Forcing authors (Feng et al., 2026) with four generation seeds per prompt, giving 512 videos per method. We report the same six VBench quality metrics as in T2V, on a 0–100 scale, and compute Avg. as their arithmetic mean. Timing uses the same H800 hardware as T2V. Training time covers one student update and four fake-score updates, and inference time covers one complete video; both are reported in seconds. The DMD2 timing entry uses the DMD2-style AR configuration.

Image-to-video generation. *Training.* We use Wan2.1-Fun-1.3B-InP-Diffusers (Alibaba PAI, 2025) for image-to-video generation. Training uses DFD's ViPE-Long data, reserving an additional held-out test split excluded from training. Both the student and fake-score model use AdamW with learning rate 10^{-5}, betas $(0.9, 0.999)$, and weight decay 0.01. The global batch size is 16, with one student update for every four fake-score updates; EMA is disabled. The backbone receives the full text condition, first-frame image representation, and corresponding mask, which remain intact during frequency decomposition. We train the DMD2 baseline for 5,000 iterations on ViPE-Long, starting its distillation from the pretrained base model rather than an existing distilled checkpoint. DP-DMD and DFD independently start from the resulting DMD2 checkpoint and train for 500 and 100 additional iterations, respectively. Our method then starts from the DFD checkpoint and continues training for another 100 iterations. At each student update of our method, the teacher uses a DMD generated-data query or a DFD real-data query, each with probability 0.5; the fake-score model always queries generated data.

Evaluation. We use 348 VBench image–prompt pairs with four generation seeds each, yielding 1,392 videos per method. We report the six VBench quality metrics used in T2V and AR and separately measure first-frame consistency through input-image subject and background preservation (I2V-S/I2V-B). All scores use a 0–100 scale, with higher values indicating better performance. Avg. includes only the six video-quality metrics, including imaging quality, and excludes I2V-S/I2V-B.

Timing uses the same H800 hardware as T2V and is reported in seconds. Training time covers one student update and four fake-score updates; inference time covers generation of one complete video. The DMD2 and DP-DMD timing configurations include GAN training, with discriminator computation included in the auxiliary updates. DMD2 takes 13.467 seconds for the student update and 36.814 seconds for the four auxiliary updates, with a reported total of 50.281 seconds; DP-

DMD takes 36.789 and 48.550 seconds, totaling 85.339 seconds. Their inference times are 15.855 and 15.593 seconds, respectively. Component and total timings retain their reported precision.

C Spatial-Frequency Diagnostics

C.1 Spatial Spectral Evaluation

Figure 1 compares our T2V model at step 40 with the official DFD step-40 checkpoint. We generate videos for the prompt sets released by the DFD and One-Forcing authors, each containing 71 prompts, using eight generation seeds per prompt. This yields 1,136 videos per model and 2,272 videos in total, forming 1,136 matched prompt–seed pairs across Ours and DFD. We analyze frame 40 from every video. Reference frames come from corresponding reference videos decoded from saved clean latents. Reference and generated frames are not pixel-aligned. This checkpoint comparison does not match additional training budgets.

We convert RGB to Rec.709 luminance, $Y = 0.2126R + 0.7152G + 0.0722B$, subtract each frame's spatial mean, and apply a two-dimensional Hann window w. Using an orthonormal spatial FFT, we compute:

$$P_f(\omega) = \frac{|\mathcal{F}_{HW}[w \odot (Y_f - \overline{Y}_f)](\omega)|^2}{\sum_{h,w} w(h,w)^2}. \tag{28}$$

We exclude DC and sum power within 64 radial bins. Frequency is normalized by the axial Nyquist frequency and extends to the diagonal limit $\sqrt{2}$. Let $P_i^m(k)$ and $P_i^{\mathrm{ref}}(k)$ denote the bin powers for condition m and its corresponding reference. We compute:

$$e_i^m(k) = \left|\log_2 \frac{\max(P_i^m(k), \epsilon)}{\max(P_i^{\mathrm{ref}}(k), \epsilon)}\right|, \qquad \bar{e}^m(k) = \frac{1}{N}\sum_{i=1}^{N} e_i^m(k), \qquad E^m = \frac{1}{64}\sum_{k=1}^{64} \bar{e}^m(k), \tag{29}$$

where $\epsilon = 10^{-12}$ and $N = 1{,}136$ matched prompt–seed cases per condition. Errors are computed for each paired case before averaging; spectra are not normalized into probability distributions. The mean errors are 1.7295 for DFD and 1.2988 for Ours. High- and low-frequency amplitude interventions yield 1.8803 and 3.3350; high- and low-frequency phase interventions yield 1.8259 and 3.8599.

C.2 Controlled Latent Amplitude and Phase Diagnostic

All four interventions operate directly on the saved complete generated video latent from Ours, without image re-encoding. We decode each modified latent with the same VAE and compare frame 40 with the unperturbed decoding. Figure 1 (right) displays Cat with a shared spatial crop. The Gaussian masks use $\sigma_f = 0.35$ and the same spatial-frequency convention as the training operators.

Write $F = \mathcal{F}_{HW}(z)$, $z_L = \mathcal{F}_{HW}^{-1}(M_L F)$, and $z_H = \mathcal{F}_{HW}^{-1}(M_H F)$. The interventions are:

$$\begin{aligned} \text{Amplitude:} \quad & z'_H = z - \alpha z_H, \quad z'_L = z - \alpha(z_L - \mu_L), \\ \text{Phase:} \quad & F'_H = F e^{i\theta M_H}, \quad F'_L = F e^{i\theta M_L}, \end{aligned} \tag{30}$$

where $\mu_L = \mathrm{mean}_{HW}(z_L)$. We set $\alpha = 0.75$, subtracting 75% of the selected soft component, and use a nominal phase bound of 180°. Phase offsets respect conjugate symmetry, preserve DC, and are modulated by the corresponding mask. Phase rotation preserves latent Fourier magnitudes and energy, but nonlinear decoding can change RGB power spectra. The unperturbed Ours decoding is the baseline, not a re-encoded image reconstruction.

D Ablation Studies

Generation seeds provide multiple outputs per prompt and are not independent training runs. All configurations within the band-specific supervision and bandwidth studies use the same training seed.

D.1 SOFT-OVERLAP LOSS WEIGHTING

The SO ablation compares training with and without Soft-Overlap Loss Weighting at 40, 60, 80, and 100 continued-training steps. Each configuration is evaluated on the prompt sets released by the DFD and One-Forcing authors, with 71 prompts per cohort and eight generation seeds per prompt, yielding 1,136 videos per configuration. Table 4 reports the six quality metrics; bold values identify the higher score within each matched-step pair. Three decimal places are retained to show small differences. This comparison studies overlap weighting and is separate from the band-specific supervision experiment below.

D.2 BAND-SPECIFIC SUPERVISION

Losses and controls. This experiment compares $\mathcal{L}_L$, $\mathcal{L}_H$, and $\mathcal{L}_L+\mathcal{L}_H$ on the same native student output head, using the non-learnable Gaussian operators with $\sigma_f = 0.35$. Each retained term has coefficient one. SO, GAN, and EMA are disabled; no full-prediction term or compensating rescaling is added. All groups use the shared complete-prediction normalizer, detached pseudo-targets, and teacher/fake-score query procedure described in Appendix B.1. Only the student supervision is changed; the fake-score model retains its original training objective.

Training. All three configurations start from the official Wan2.1-T2V-1.3B DFD checkpoint at iteration 40 and continue for 40 iterations, comprising eight student updates and 32 fake-score updates. Training uses eight GPUs with FSDP, global batch size 16, per-GPU microbatch size one, two accumulation rounds, and learning rate 10^{-5} for both models. Each fake-score model starts from the original Wan teacher, and optimizers are initialized afresh. The training split contains 28,986 ViPE-Long latent examples, with prompt-overlap checks against the evaluation prompts. Model computation uses BF16, with FP32 filtering and residual spectral energy. The teacher remains frozen and uses CFG 5, with real-data query probability 0.5.

Generation and scope. Evaluation uses 71 prompts from ViPE-Long, with eight generation seeds per prompt. Each group therefore has 568 videos, giving 1,704 videos across the three configurations. We use four-step DFD sampling with times $[0.999, 0.937, 0.833, 0.624, 0]$ and non-EMA student weights. No additional classifier-free guidance is applied during student inference (guidance scale 1); the teacher guidance scale used during training is 5. Videos contain 81 frames at 832×480 and 16 fps.

Quantitative results. Table 6 compares Low-only, High-only, and Low+High supervision. The six VBench scores are displayed on a 0–100 scale. Since all groups start from DFD40, the labels describe only the continued-training supervision. Removing a loss term changes both the supervised frequencies and the total gradient scale.

Table 6: Band-only supervision ablation with $\sigma_f = 0.35$ and SO disabled. All metrics are higher-is-better; bold marks the highest score in each column.

Loss combination	Subject Consist.	Backgr. Consist.	Temporal Flicker.	Motion Smooth.	Aesthetic Quality	Imaging Quality
Low-only	95.8728	95.4641	97.7333	98.7653	63.8894	70.8663
High-only	95.7628	95.3337	**97.7936**	98.7744	63.6015	70.5548
Low+High	**95.9216**	**95.5006**	97.7830	**98.7865**	**63.9001**	**70.8702**

D.3 FREQUENCY BANDWIDTH

Controlled comparison. We compare $\sigma_f \in \{0.1, 0.35, 0.6\}$ in our method, using the normalized spatial frequency coordinates and complementary Gaussian masks in Equation 1. Only the bandwidth changes. All arms use equally weighted low- and high-frequency losses, a shared complete-prediction normalizer, and no additional full-prediction loss. SO is disabled by setting both overlap exponents to zero. The filters act on the spatial dimensions of the same complete student prediction;

no separate frequency heads are introduced. Increasing σ_f raises the low-pass mask value at each nonzero frequency and correspondingly lowers the high-pass value.

Training and evaluation. Each arm starts from official DFD checkpoint 40 and trains for 40 additional outer iterations, comprising eight student updates and 32 fake-score updates. Training uses eight GPUs, global batch size 16, and learning rate 10^{-5}. The checkpoint at continued iteration 40 is evaluated for each bandwidth. The supplied protocol comparison confirms matching fields apart from the bandwidth and entry-file fingerprints, with identical shared training-code fingerprints.

Evaluation uses text-only conditioning on 71 prompts from ViPE-Long, with eight generation seeds per prompt, giving 568 videos per configuration. Inference uses the native complete prediction with four-step DFD sampling, producing 81 frames at 832×480 and 16 fps. Table 7 reports the six VBench dimensions multiplied by 100.

Table 7: Gaussian bandwidth ablation with SO disabled. All metrics are higher-is-better; bold marks the highest score in each column.

σ_f	Subject Consist.	Backgr. Consist.	Temporal Flicker.	Motion Smooth.	Aesthetic Quality	Imaging Quality
0.1	95.8135	95.4418	**97.8022**	**98.7952**	63.7338	70.4771
0.35	**95.9216**	95.5006	97.7830	98.7865	63.9001	**70.8702**
0.6	95.8963	**95.5111**	97.7622	98.7791	**63.9102**	70.7792

Interpretation. Table 7 shows tradeoffs between visual quality, consistency, and temporal stability across bandwidths, rather than uniform improvements as bandwidth increases. We choose $\sigma_f = 0.35$ as a favorable balance among these aspects.

E Additional Qualitative Results

E.1 Text-to-Video Generation

Figure 6 presents additional comparisons for a desert train and a futuristic city, complementing the examples in Section 5.1.

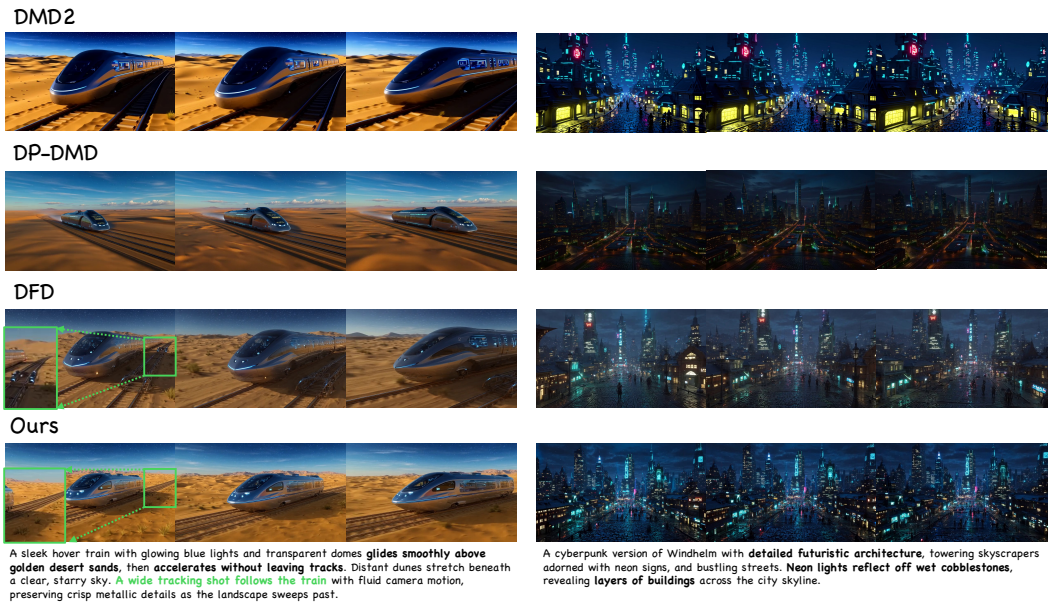


Figure 6: Additional text-to-video comparisons for a desert train and a futuristic city.

E.2 Image-to-Video Generation

Figure 7 supplements the image-conditioned results in Section 5.3.

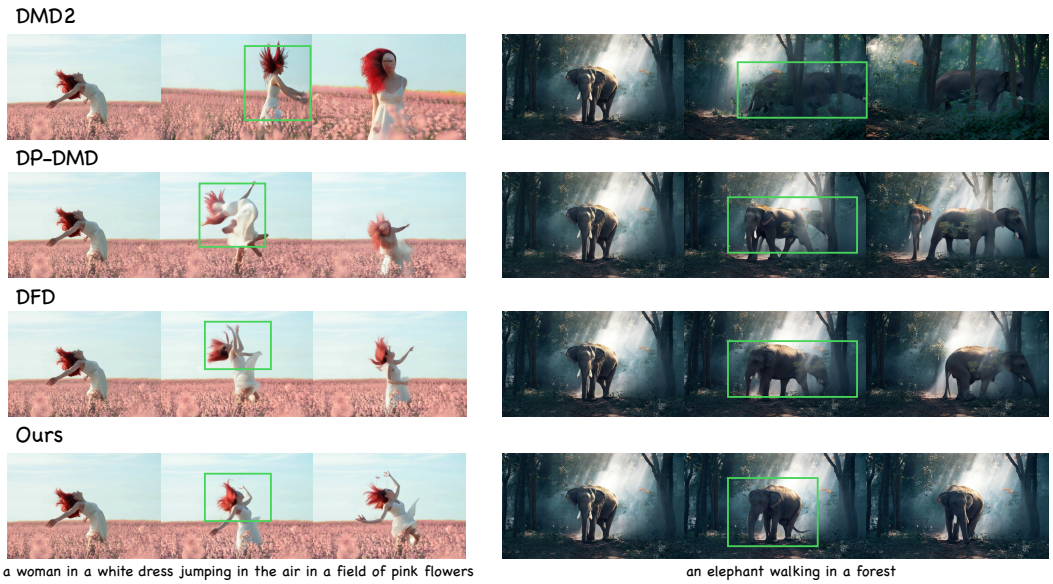


Figure 7: Image-to-video comparisons for a woman in a flower field and an elephant in a forest.