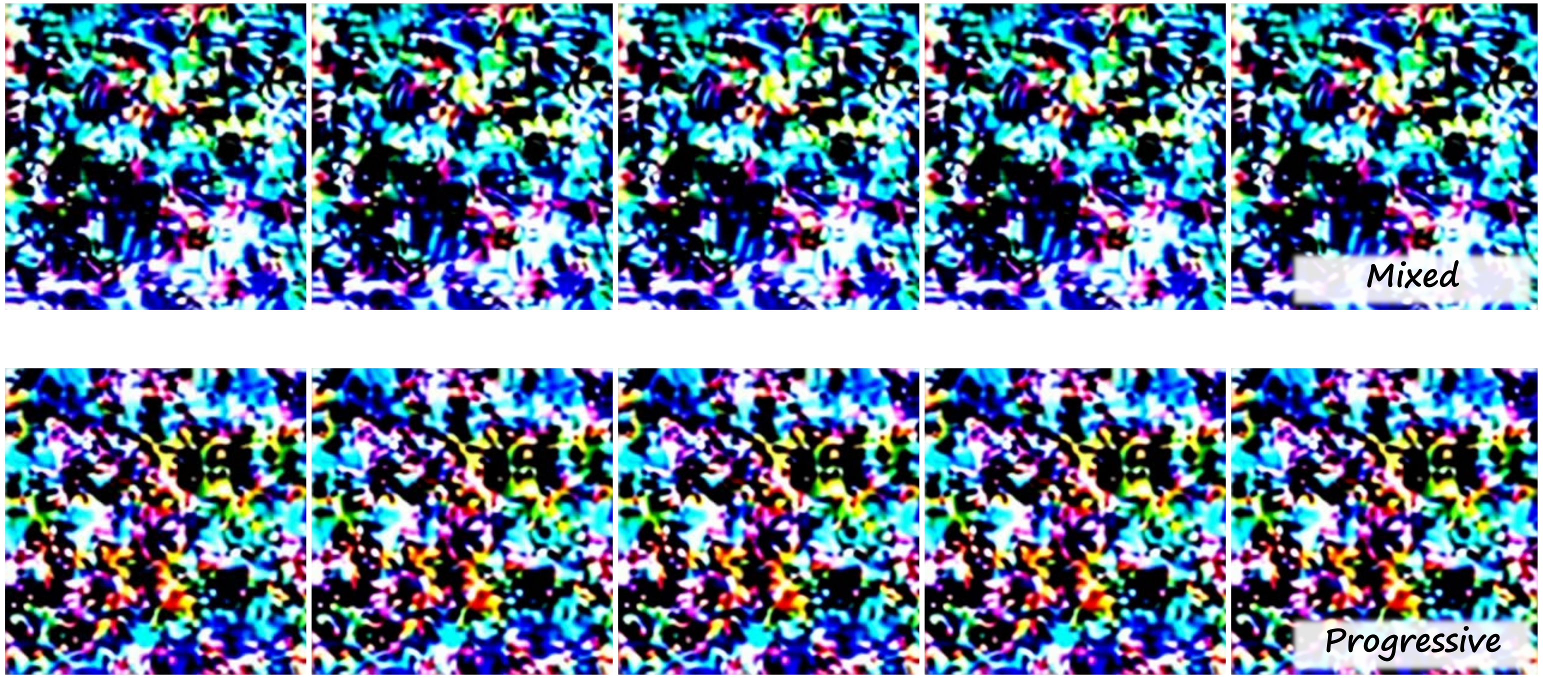
FreqPrior: Improving Video Diffusion Models with Frequency Filtering Gaussian Noise

Yunlong Yuan[1], Yuanfan Guo[2], Chunwei Wang[2], Wei Zhang[2], Hang Xu[2], Li Zhang[1]

[1]School of Data Science, Fudan University
[2]Noah's Ark Lab, Huawei

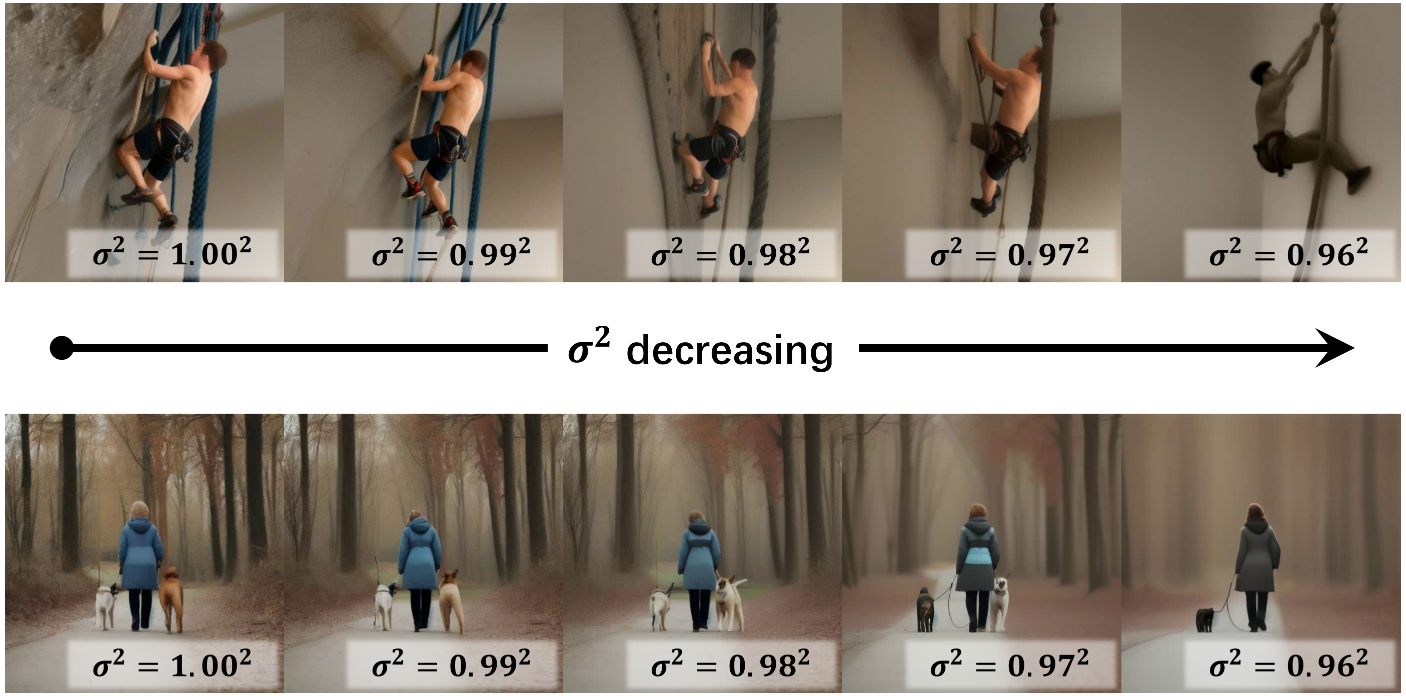
Noise priors of video diffusion models

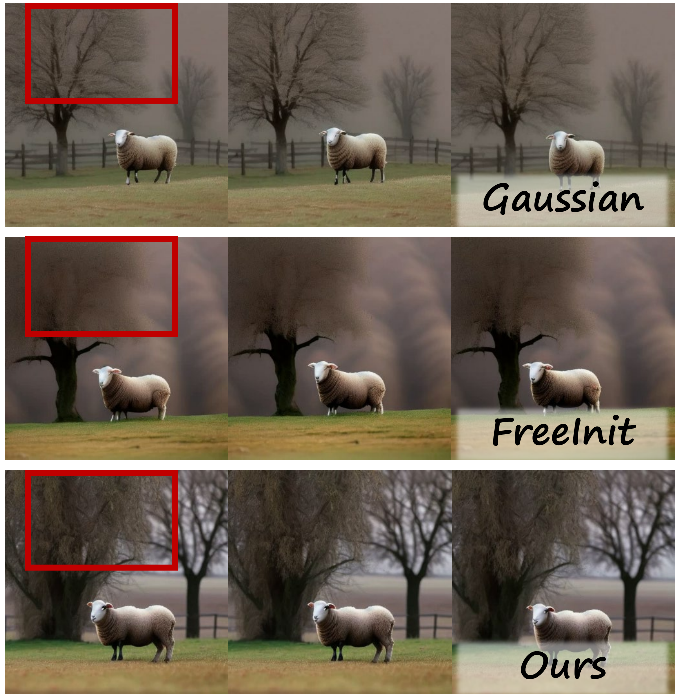


Noise priors proposed by PYoCo [Ge et al. 2023] establish temporal correlations across frames.

However, they require extensive training and cannot be applied to other pretrained video diffusion models directly.

Noise priors of video diffusion models

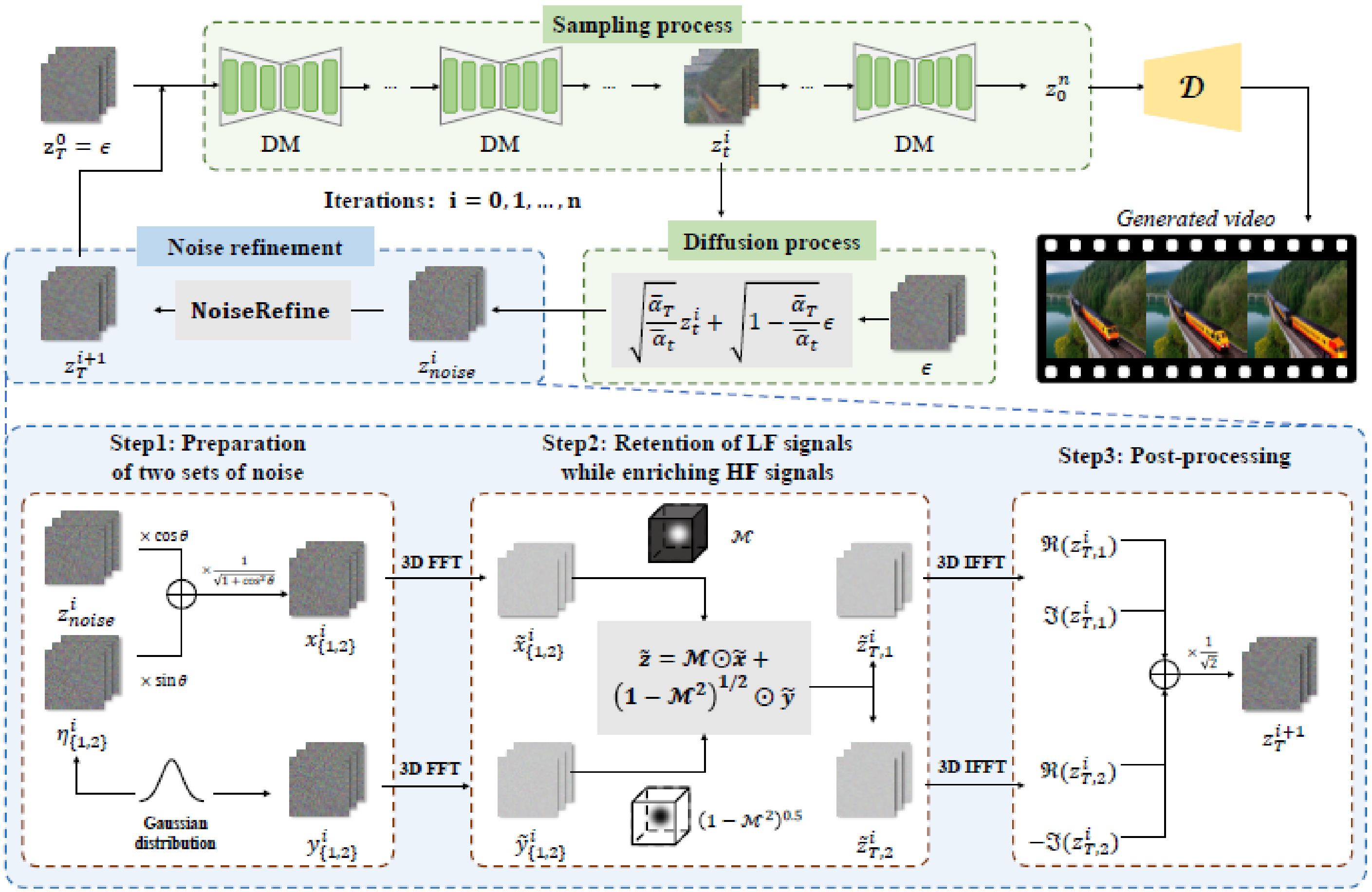




FreeInit [Wu et al. 2024] uses classical frequency filtering on the noise prior to enhance the temporal consistency, but the generated videos suffer from excessive smoothness, limited motion dynamics, and a lack of details.

Pipeline of FreqPrior

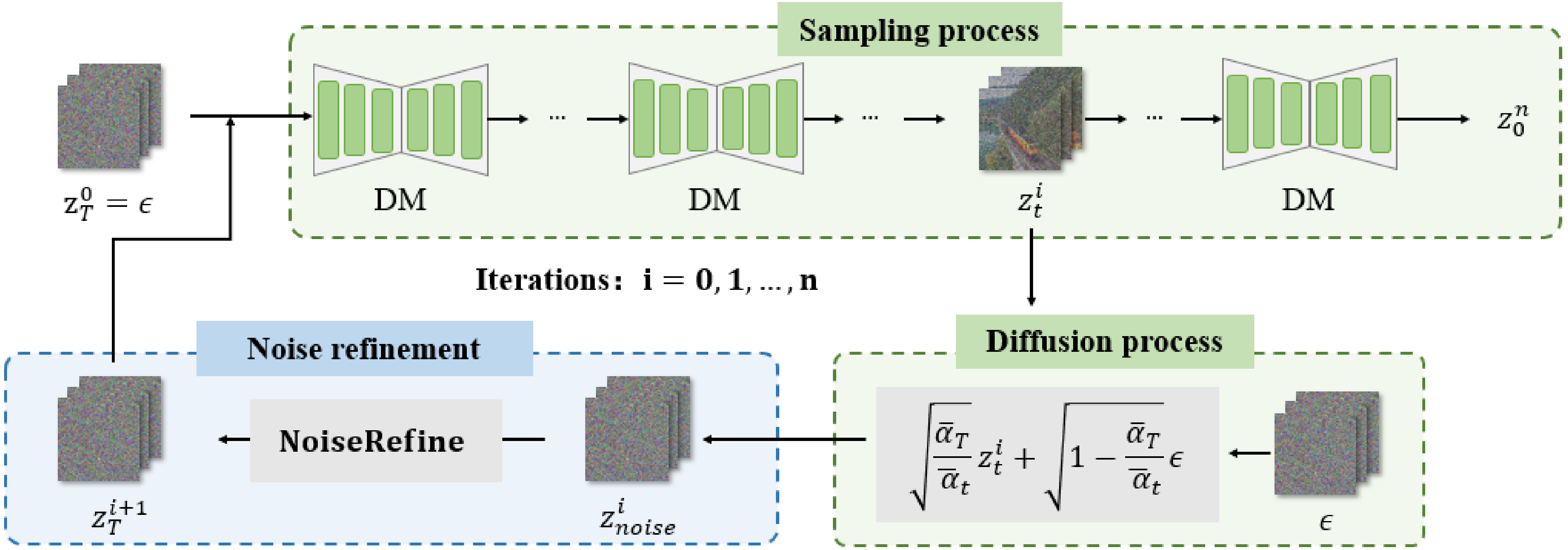
- sampling process
- diffusion process
- noise refinement



Sampling process and diffusion process

Sampling process: DDIM sampling [Song et al. 2021].

Diffusion process: diffuses the latent with initial noise ϵ at an intermediate timestep.



Noise refinement

- **Preparation of two sets of noise**

We prepare two distinct sets of noise. One set is to convey low-frequency information:

$$x_1^i = \frac{1}{\sqrt{1+\cos^2\theta}}\left(\cos\theta\cdot z_{noise}^i + \sin\theta\cdot\eta_1^i\right), \qquad \eta_1^i \sim \mathcal{N}(0, I),$$
$$x_2^i = \frac{1}{\sqrt{1+\cos^2\theta}}\left(\cos\theta\cdot z_{noise}^i + \sin\theta\cdot\eta_2^i\right), \qquad \eta_2^i \sim \mathcal{N}(0, I).$$

The other set of noise is designed to provide high-frequency details: y_1^i and y_2^i are independent Gaussian noise.

- **Frequency filtering**

We map the noise to the frequency domain:

$$\tilde{x}_{\{1,2\}}^i = \mathcal{F}_{3D}\left(x_{\{1,2\}}^i\right), \qquad \tilde{y}_{\{1,2\}}^i = \mathcal{F}_{3D}\left(y_{\{1,2\}}^i\right).$$

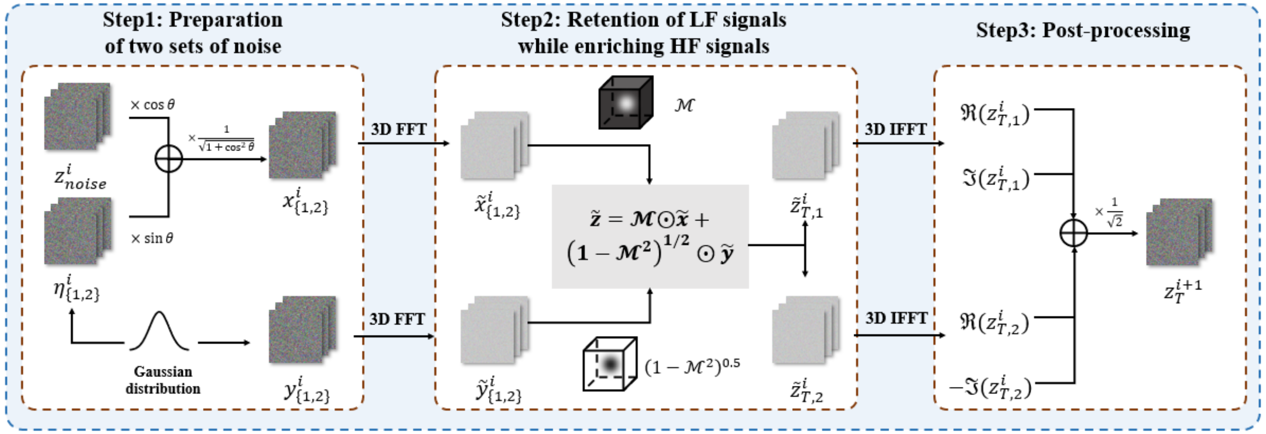
Then, we perform frequency filtering.

$$\tilde{z}_1^i = \mathcal{M}\odot\tilde{x}_1^i + (1-\mathcal{M}^2)^{0.5}\odot\tilde{y}_1^i, \qquad \tilde{z}_2^i = \mathcal{M}\odot\tilde{x}_2^i + (1-\mathcal{M}^2)^{0.5}\odot\tilde{y}_2^i.$$

- **Post-processing**

After filtering, the frequency features are mapped back into the latent space, followed by post-processing to form the new noise prior z_T^{i+1}:

$$z_T^{i+1} = \frac{1}{\sqrt{2}}\left(\Re(z_{T,1}^i) + \Im(z_{T,1}^i) + \Re(z_{T,2}^i) - \Im(z_{T,2}^i)\right), \qquad z_{T,\{1,2\}}^i = \mathcal{F}_{3D}^{-1}\left(\tilde{z}_{\{1,2\}}^i\right).$$



Theoretical analysis

Assumption *After the diffusion process,* z_{noise} *follows a standard Gaussian distribution* $\mathcal{N}(0, I)$.

Theorem *Given a DFT matrix or multi-dimension DFT matrix* $F \in C^{N \times N}$, *with* A *and* B *are its real part and imaginary part respectively, it holds that* $AB = BA = 0$ *and* $A^2 + B^2 = NI$.

Considering frame, height and width dimensions, we can infer the distribution of FreeInit [Wu et al. 2024] noise prior and our method:

FreeInit: $$z \sim \mathcal{N}(0, (I - P)^2 + P^2), \qquad P = \frac{1}{N}(A\Lambda A + B\Lambda B).$$

FreqPrior: $$z \sim \mathcal{N}\left(0, I - \frac{2\cos^2\theta}{1 + \cos^2\theta} Q^2\right), \qquad Q = \frac{1}{N}(A\Lambda B + B\Lambda A).$$

Theoretical results and numerical results

Theoretical results Under the condition of the same low-pass filter $\mathcal{M}$, we can derive:

$$\left\|I - \Sigma_{FreqPrior}\right\|_F \leq \frac{\cos^2\theta}{1+\cos^2\theta}\|I - \Sigma_{FreeInit}\|_F.$$

Numerical results

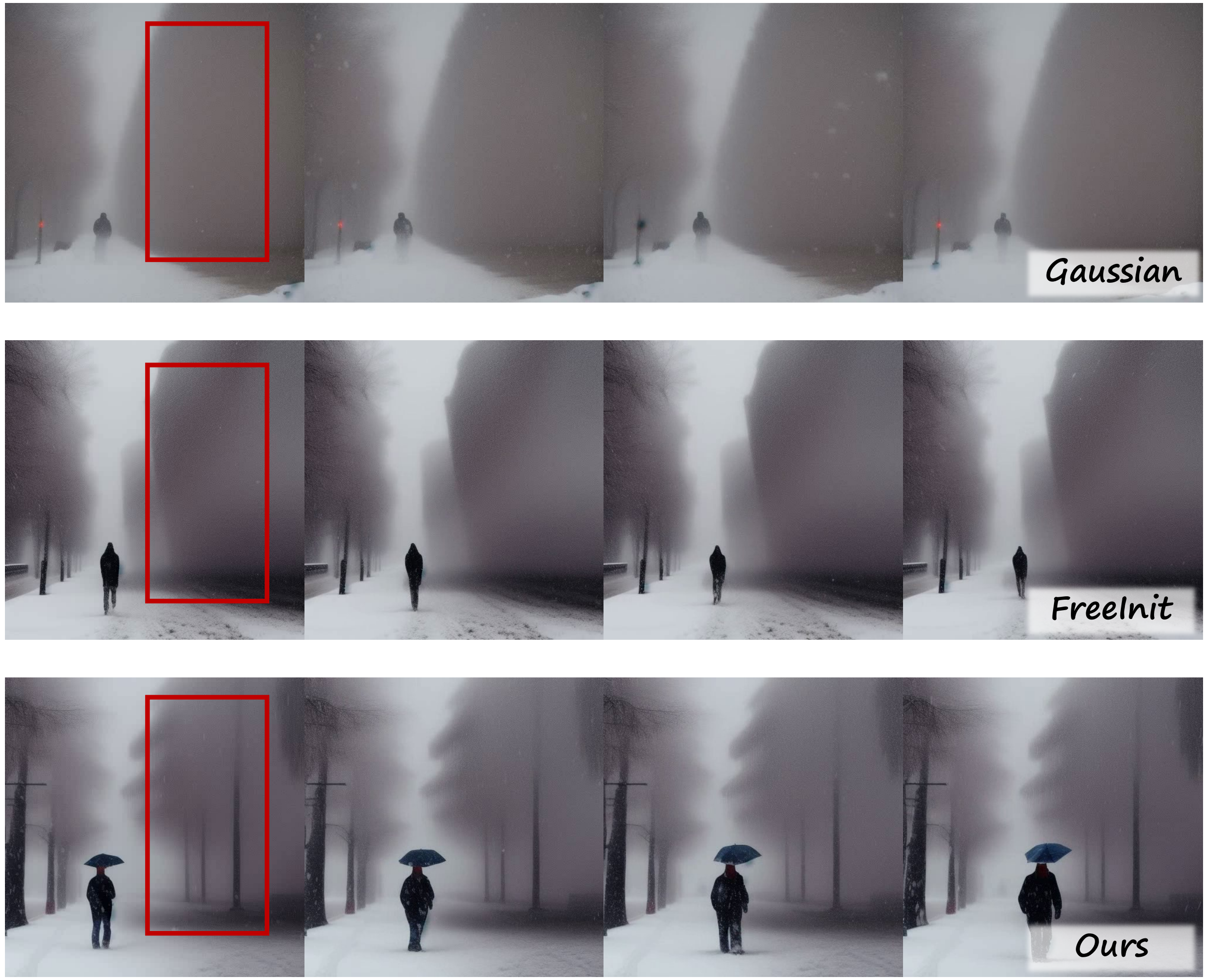
Prior	(16, 20, 20)		(16, 30, 30)		(16, 40, 40)	
	Butterworth	Gaussian	Butterworth	Gaussian	Butterworth	Gaussian
Mixed	154.9193		232.3790		309.8387	
FreeInit	3.8230	8.5878	5.7001	12.8817	7.6026	17.1756
Ours	8.5071×10^{-28}	7.7218×10^{-28}	1.4002×10^{-26}	1.2656×10^{-26}	2.7342×10^{-26}	2.4140×10^{-26}

Experimental results

Base model	Noise prior	Prior finding	Generation	Quality	Semantic	Total	Inference time
VideoCrafter	Gaussian	/	25 steps	69.50	54.92	66.58	**27.73s**
	Mixed	/	25 steps	–	–	–	–
	Progressive	/	25 steps	–	–	–	–
	Gaussian	/	3*25 steps	69.75	58.10	67.42	83.09s
	FreeInit	2 full sampling	25 steps	70.62	58.97	68.29	83.18s
	Ours	2 partial sampling	25 steps	**70.63**	**61.33**	**68.77**	63.67s
ModelScope	Gaussian	/	50 steps	73.13	65.69	71.64	**19.24s**
	Mixed	/	50 steps	–	–	–	–
	Progressive	/	50 steps	–	–	–	–
	Gaussian	/	3*50 steps	73.25	66.31	71.87	57.72s
	FreeInit	2 full sampling	50 steps	73.61	67.24	72.34	57.73s
	Ours	2 partial sampling	50 steps	**74.04**	**69.06**	**73.04**	44.88s
AnimateDiff	Gaussian	/	25 steps	79.56	69.03	77.45	**23.34s**
	Mixed	/	25 steps	–	–	–	–
	Progressive	/	25 steps	–	–	–	–
	Gaussian	/	3*25 steps	79.49	69.71	77.54	70.22s
	FreeInit	2 full sampling	25 steps	79.58	68.85	77.43	70.45s
	Ours	2 partial sampling	25 steps	**80.05**	**70.37**	**78.11**	54.05s

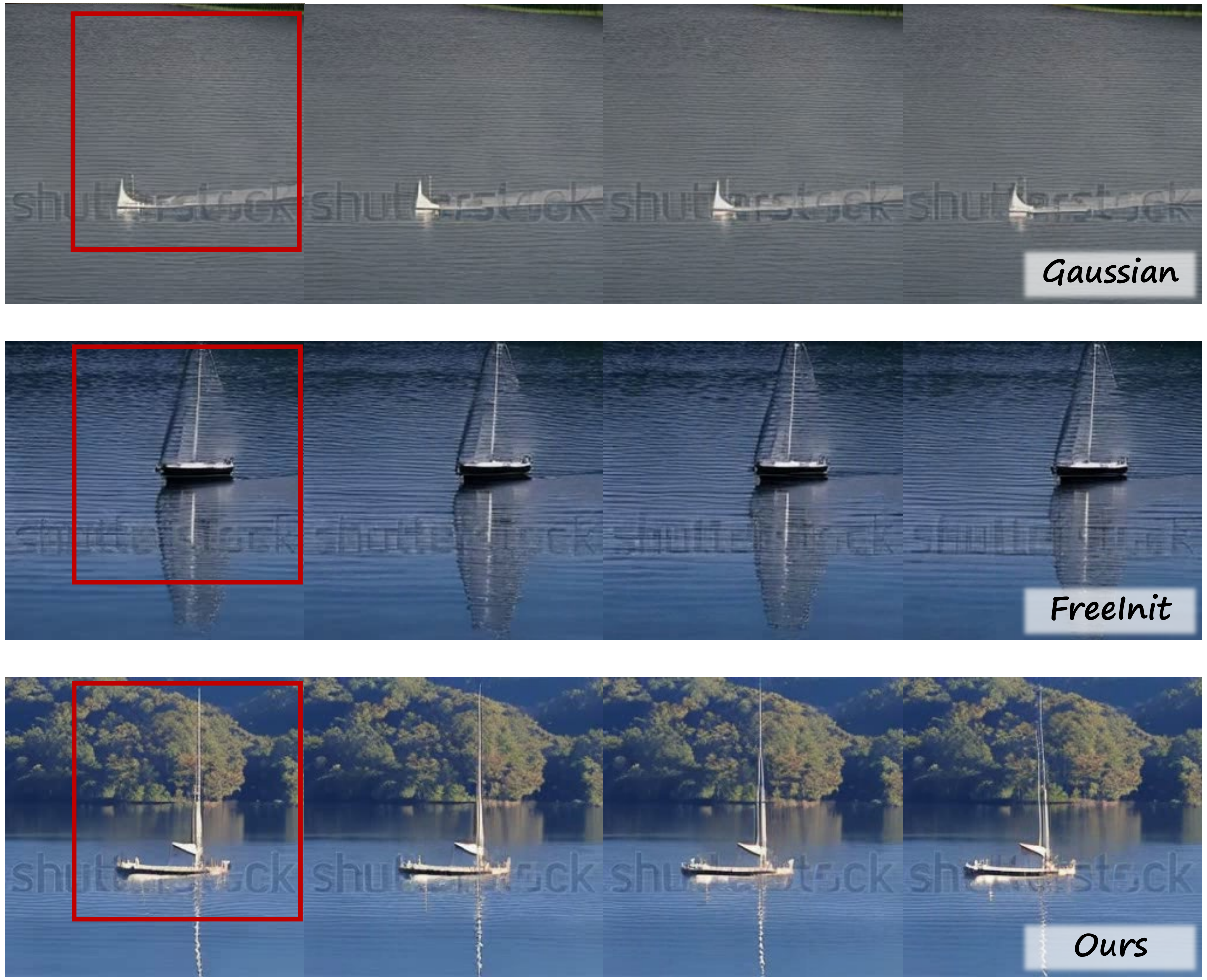
Qualitative comparisons

a person drinking coffee in a **cafe**



a person walking in the snowstorm

Qualitative comparisons

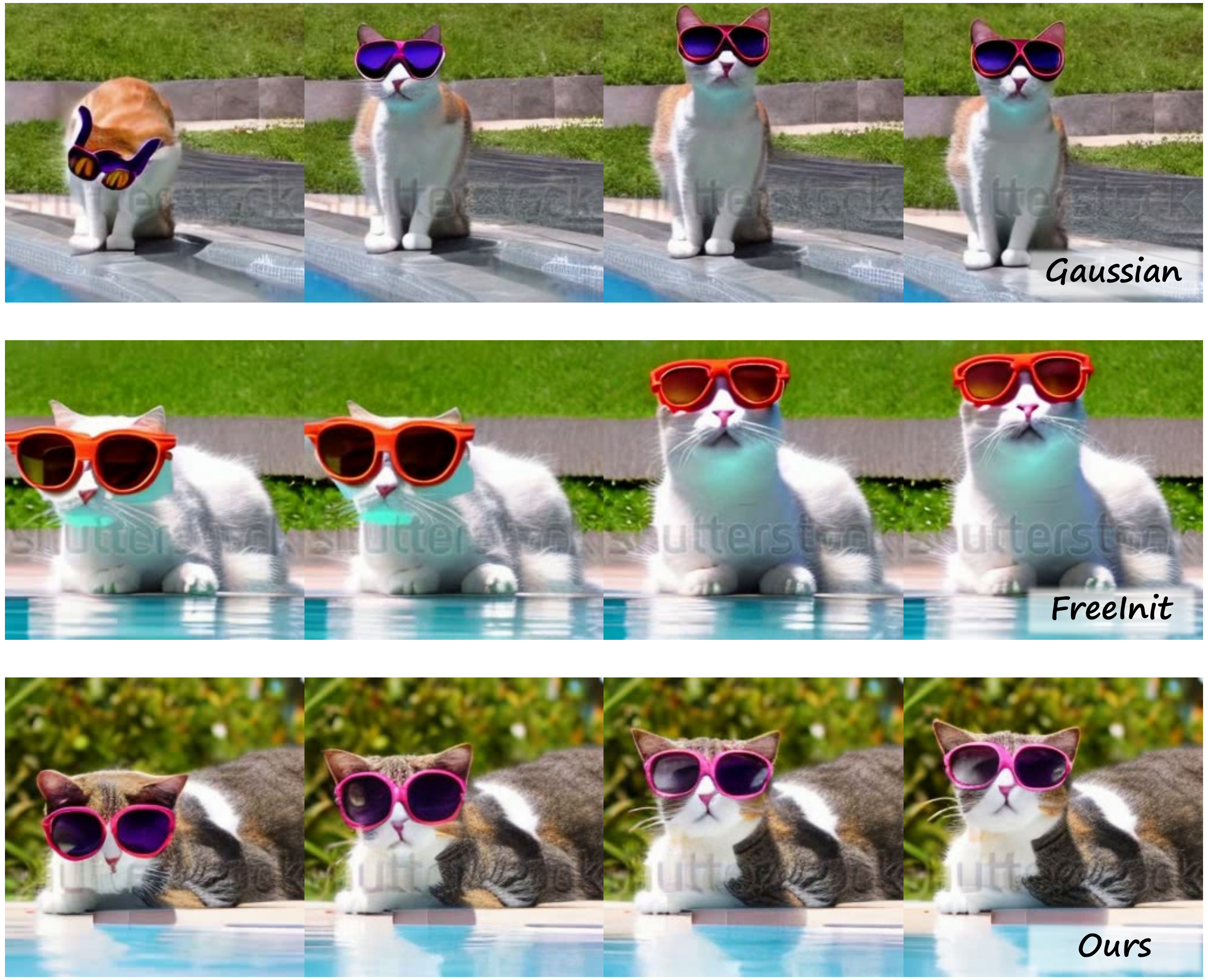


a boat sailing smoothly on a calm lake

A happy fuzzy panda playing guitar nearby a campfire, snow mountain in the background

Qualitative comparisons

An oil painting of a couple in formal evening wear going home get caught in a heavy downpour with umbrellas



A cat wearing sunglasses at a pool

ICLR

Thank you!

FreqPrior: Improving Video Diffusion Models with Frequency Filtering Gaussian Noise

Project page:
https://github.com/fudan-zvg/FreqPrior