FreeInit: Bridging Initialization Gap in Video Diffusion Models

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu✉

S-Lab, Nanyang Technological University
{tianxing001, chenyang.si, yuming002, ziqi002, ziwei.liu}@ntu.edu.sg
https://tianxingwu.github.io/pages/FreeInit/

Abstract. Though diffusion-based video generation has witnessed rapid progress, the inference results of existing models still exhibit unsatisfactory temporal consistency and unnatural dynamics. In this paper, we delve deep into the noise initialization of video diffusion models, and discover an implicit training-inference gap that attributes to the unsatisfactory inference quality. Our key findings are: **1)** the spatial-temporal frequency distribution of the initial noise at inference is intrinsically different from that for training, and **2)** the denoising process is significantly influenced by the low-frequency components of the initial noise. Motivated by these observations, we propose a concise yet effective inference sampling strategy, **FreeInit**, which significantly improves temporal consistency of videos generated by diffusion models. Through iteratively refining the spatial-temporal low-frequency components of the initial latent during inference, FreeInit is able to compensate the initialization gap between training and inference, thus effectively improving the subject appearance and temporal consistency of generation results. Extensive experiments demonstrate that FreeInit consistently enhances the generation results of various text-to-video generation models without additional training.

Keywords: Video diffusion models · Initial noise · Temporal consistency

1 Introduction

Recently, diffusion models have demonstrated impressive generative capabilities in text-to-image generation [31, 32, 35]. These advancements have attracted substantial attention, highlighting the potential of creating diverse and realistic images based on textual descriptions. In light of these achievements, researchers are now exploring the application of diffusion models in text-to-video (T2V) generation [1, 4, 10, 12, 13, 36, 45, 46, 50, 51], with the goal of synthesizing visually appealing and contextually coherent videos from textual descriptions. Most of these video diffusion models are built upon powerful pretrained image diffusion models, *e.g.*, Stable Diffusion (SD) [32]. Through the incorporation of temporal layers and large-scale training on extensive video datasets, these models are capable of generating video clips that align with the given text prompts. Despite

Fig. 1: FreeInit for Video Generation. We propose ***FreeInit***, a concise yet effective method to significantly improve temporal consistency of videos generated by diffusion models. FreeInit requires no additional training and introduces no learnable parameters, and can be easily incorporated into arbitrary video diffusion models at inference time.

these advancements, the videos generated by these models often suffer from issues related to temporal inconsistency and unnatural dynamics.

In this paper, we delve into the impact of noise initialization on video generation, identifying a significant disparity between the training and the inference process. Specifically, we find that the diffusion process fails to fully corrupt the clean latent into pure Gaussian noise, especially in the low-frequency band. To illustrate this, Figure 2 shows the frames decoded from noisy latent during the diffusion process, alongside a spatio-temporal frequency decomposition to assess the extent of corruption across different frequency bands. Remarkably, the corruption of low-frequency components occurs at a notably slower rate than that of the high-frequency components. As a result, the noisy latent at the final diffusion step (t=1000) will still contain considerable low-frequency information from the input video. Since the real video frames are temporally correlated in nature, this information leakage eventually leads to an implicit gap between training and inference: at training, the initial noises corrupted from real videos remain temporally correlated at low-frequency band, while during inference, the i.i.d Gaussian initial noise is entirely uncorrelated. Furthermore, we discover that these low-frequency components can substantially impact the quality of the generated videos, as revealed in our observations in Figs. 5 and 6. Thus, when applying the diffusion models trained with the correlated initial noises to non-correlated Gaussian initial noise at inference, the performance deteriorates, exhibiting unsatisfactory temporal consistency and unnatural motions.

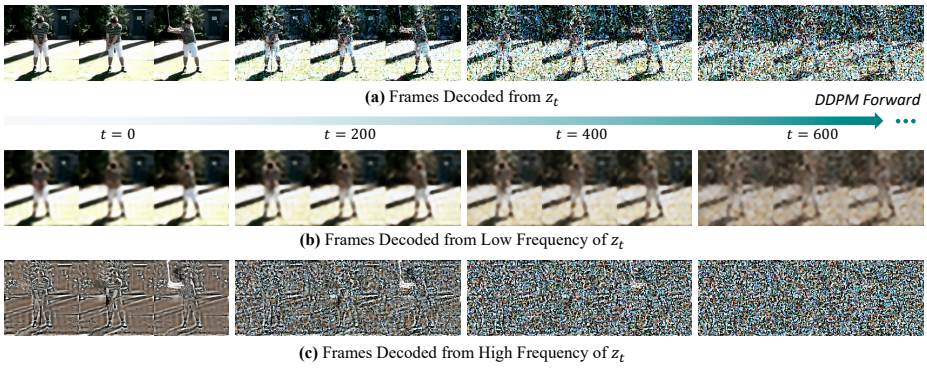


Fig. 2: Visualization of Decoded Noisy Latent from Different Spatio-Temporal Frequency Bands at Training. (a) Video frames decoded from the entire frequency band of the noisy latent z_t in DDPM Forward Process. (b) Frames decoded from the low-frequency components of z_t. It is evident that the diffusion process has difficulty in fully corrupting the semantics, leaving substantial spatio-temporal correlations in the low-frequency components. (c) Frames decoded from the high-frequency components of z_t. Each frame degenerates rapidly with the diffusion process.

Motivated by these observations, we propose a novel inference-time sampling method, denoted as ***FreeInit***, to bridge the initialization gap between training and inference without any additional training or fine-tuning. Specifically, during the inference process, we first initialize an independent Gaussian noise, which then undergoes the DDIM denoising process to generate a clean video latent. Subsequently, we obtain a noisy version of the clean video latent through the forward diffusion process. Since this noisy latent is obtained from the denoised latent rather than pure noise, its low-frequency components have improved temporal consistency. With this noisy latent, we proceed to reinitialize the noise by combining its low-frequency components with the high-frequency components from a random Gaussian noise using spatio-temporal frequency filter. Finally, this reinitialized noise serves as the starting point for a new round of DDIM sampling. By iterating this refinement process several times, the initial noise at inference is gradually guided towards the training distribution, facilitating the generation of frames with enhanced temporal consistency and visual appearance.

Extensive experiments across diverse evaluation prompt sets demonstrate the steady enhancement brought about by FreeInit for various text-to-video generation models. As illustrated in Fig. 1, FreeInit plays a significant role in improving temporal consistency and the visual appearance of generated frames. This method can be readily applied during inference without the need for parameter tuning. Furthermore, to achieve superior generation quality, the frequency filter can be conveniently adjusted for each customized base model. We summarize our contributions as follows:

– We systematically investigate the noise initialization of video diffusion models, and identify an implicit training-inference gap that contributes to the

inference quality drop. To our knowledge, we are the first to study the impact of initial noise on video diffusion models from the frequency domain.
- We propose a concise yet effective sampling strategy, referred to as FreeInit, which iteratively refines the initial noise without the need for additional training or fine-tuning.
- Extensive quantitative and qualitative experiments demonstrate that FreeInit can be effectively applied to various text-to-video models. It consistently improves the inference quality of generated videos.

2 Related Work

Video Generative Models. There are mainly three types of video generation models, namely GAN-based [8], transformer-based [43], and diffusion-based [14]. StyleGAN-V [37], MoCoGAN-HD [42], and [2] utilize the powerful StyleGAN [19–21] to generate videos. Transformer-based models [16,18,44,47,48] such as Phenaki [44], CogVideo [16], and NÜWA [48] encode videos as visual tokens and train transformer models to auto-regressively generate the visual tokens. Recently, diffusion models [5, 14, 38, 40] have made remarkable progress in text-to-image generation [26, 28, 32, 35], and have enabled a line of works that extends these pre-trained diffusion models towards text-to-video generation [1, 7, 10–13, 15, 22, 25, 36, 45, 46, 50–52]. In this work, our method is built on top of diffusion-based text-to-video methods. We propose to iteratively refine the initial noise to improve temporal consistency of pre-trained video diffusion models. We demonstrate the effectiveness of our method on various diffusion models, including VideoCrafter, ModelScopeT2V (denoted as ModelScope), and AnimateDiff. VideoCrafter [12] employs the pre-trained text-to-image model Stable Diffusion [32] and incorporates newly initialized temporal layers to enable video generation. ModelScope [45] also initializes the spatial part from Stable Diffusion and adds spatio-temporal block to learn temporal dependencies. AnimateDiff [10] trains motion modeling modules and inserts them into personalized text-to-image diffusion models to achieve animated videos of customized concepts, *e.g.*, characters, styles, *etc*.

Noise in Diffusion Models. Only a few previous works have mentioned the limitations of the noise schedule of current diffusion models. In the image domain, [23] points out common diffusion noise schedules cannot fully corrupt information in natural images, limiting the model to only generate images with medium brightness. A rescaled training schedule is then proposed to alleviate this problem through fine-tuning. Recently, [6] makes further discussions on the signal leakage issue, and propose to explicitly model the signal leakage for better inference noise distribution, which produces images with more diverse brightness and colours. A resampling operation similar to our iterative refinement strategy is proposed in [24] to harmonize the inpainted image across full inference timesteps. Different from this, we tackle the initialization problem and explore in frequency-domain to improve temporal consistency. In the video domain, PYoCo [7] carefully designs the progressive video noise prior to achieve a better

video generation performance. Similar to [23], PYoCo also focuses on the noise schedule at training stage and requires massive fine-tuning on video datasets. In contrast, we focus on the initial noise at inference stage and proposes a concise inference-time sampling strategy that bridges the training-inference discrepancy with no fine-tuning required. Some recent works [9,29] also pay attention to the inference initial noise, but aiming at generating long videos. We instead focus on improving inference quality, and further design specific frequency-domain-based operations to modulate different frequency components of the initial noise.

3 Preliminaries and Observations

3.1 Preliminaries

Similar to image diffusion models, *training* video diffusion models also involve a diffusion process and a denoising process, and operate in the latent space of an autoencoder. The diffusion process includes a sequence of T steps. At each step t, Gaussian noise is incrementally added to the video latent $\boldsymbol{z}_0$, following a predefined variance schedule $\beta_1, \ldots, \beta_T$:

$$q(z_{1:T}|z_0) = \prod_{t=1}^{T} q(z_t|z_{t-1}), \tag{1}$$

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}). \tag{2}$$

Let $\alpha_t = 1 - \beta_t$, $\overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\overline{\alpha}_t} z_0, (1-\overline{\alpha}_t)\mathbf{I}). \tag{3}$$

As a result, the noisy latent z_t at each timestep t can be directly sampled as:

$$z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \tag{4}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a Gaussian white noise with the same shape as z_t.

In the reverse process, the network learns to recover the clean latent z_0 by iterative denoising with U-Net [33], starting from the initial noise z_T:

$$p_\theta(z_{0:T}) = p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t), \tag{5}$$

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \tag{6}$$

where μ_θ and Σ_θ are predicted by the denoising U-Net ϵ_θ.

During *inference*, an initial latent $\hat{z_T}$ is first initialized, typically as a Gaussian noise sampled from normal distribution:

$$\hat{z_T} = \epsilon' \sim \mathcal{N}(0, \mathbf{I}). \tag{7}$$

Then the trained network ϵ_θ is used to iteratively denoise the noisy latent to a clean latent $\hat{z_0}$ through DDIM sampling [39], which is then decoded with decoder $\mathcal{D}$ to obtain video frames $\hat{x_0}$.

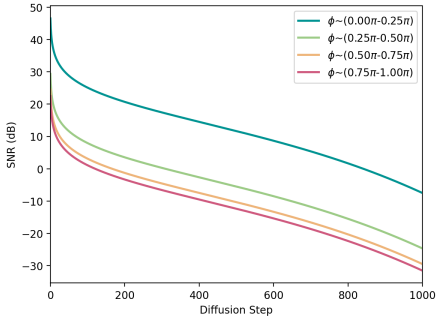


Fig. 3: Signal-to-Noise Ratio (SNR) of different frequency bands at the forward diffusion process. Each curve corresponds to a spatio-temporal frequency band of the latent code z_t when adding noise at training. The pattern indicates a much slower corruption on low-frequency components.

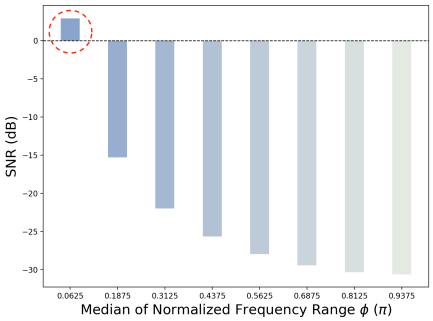


Fig. 4: Frequency Distribution of the SNR in the initial noise. When training with the typical Stable Diffusion Noise Schedule, the SNR of the initial noise is extremely high in low-frequency components, even larger than 0 dB (red circle). This indicates a severe information leak at the low-frequency band.

3.2 The Initialization Gap

Information Leakage at Training. At training stage, the network learns to denoise the corrupted latent obtained from the forward diffusion process. However, we find the commonly used diffusion strategy has difficulty in fully corrupting information from real videos, especially in their spatio-temporal low-frequency band. To better demonstrate this phenomenon, we utilize Signal-to-Noise Ratio (SNR) to measure the amount of preserved information at the forward diffusion process. Fig. 3 shows the SNR measurements of the noisy latent z_t (as defined in Eq. (4)) corrupted from a random video clip using Stable Diffusion noise schedule. The figure reveals an obvious pattern wherein the low-frequency components (blue-green curve) exhibit a significantly slower corruption rate compared to the high-frequency components (red curve), which aligns with our observation in Fig. 2. Furthermore, we analyze the average SNR distribution of the initial noises z_T (T=1000) on UCF-101, and find that the SNR in low-frequency band is even larger than 0 dB, indicating a severe leakage of low-frequency information into the initial noise (Fig. 4).

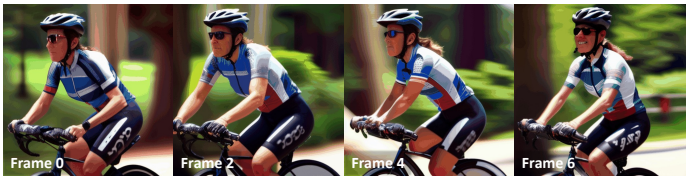
These observations demonstrate the existence of an implicit gap between the training and inference processes. Specifically, the noise introduced during training is insufficient to completely corrupt video information, causing the low-frequency components of the initial noise (*i.e.*, latent at t=1000) persistently contain spatio-temporal correlations. However, during the inference process, the video generation model is tasked with generating coherent frames from non-correlated Gaussian noise. This presents a considerable challenge for the denoising network, as its initial noise lacks spatio-temporal correlations at inference. For instance, as illustrated in Fig. 6, the "biking" video generated from Gaussian

Fig. 5: Role of Initial Low-Frequency Components. Each column shows three frames generated from the mixed initial noise. We observe that even if the majority (*e.g.*, 80%) of high frequencies are replaced, the generated results still remain largely similar to the original "Full z_T" frames, indicating that the overall distribution of the generated results is determined by the low-frequency components of the initial noise.

noise exhibits unsatisfactory temporal consistency. In contrast, when using the corrupted latent obtained through the forward diffusion process from real videos as initial noise, the generated frames showcase improved temporal consistency.

Influence of Initial Low-frequency Components. Considering the SNR gaps of initial noise between training and inference, we further investigate the influence of the low-frequency components of initial noise. A noisy latent z_T is first obtained from diffusing a real video. Then its high-frequency components are gradually removed and replaced with that of a random Gaussian noise, only keeping low-frequencies unchanged. Finally, the mixed latent is used as initial noise for inference. As shown in Fig. 5, it is evident that variations in high-frequency band have a negligible impact on the overall generation results. Remarkably, the overall distribution of the generated outcomes remains stable, even when employing only 20% of the original initial latent information from the low-frequency band. When all information is removed, the denoising process equates with pure Gaussian noise initialization, which leads to relatively poor generation results. This observation highlights two key conclusions: 1) the low-frequency components of the initial noise play a dominant role at inference, and 2) the quality of low-frequency components is crucial for the generation quality. Our hypothesis is that this is due to the aforementioned information leak during training, which biases the denoising towards the low-frequency components of initial noise. These conclusions motivate us to propose a concise yet effective strategy for enhancing the inference quality of video diffusion models.

(a) Inference with Gaussian Noise



(b) Inference with Training Initial Noise

Fig. 6: Initialization Gap. (a) With randomly initialized Gaussian noise for different frames, the sampled video exhibits inconsistency among frames. (b) When we start from noisy latent obtained from the diffusion process from real videos, the generated video is temporally consistent. This is because the initial noise is aligned with training stage and it contains correlated information among different frames in nature.

4 FreeInit

Motivated by the above analysis, we propose a method for relieving this gap by progressively refining the low-frequency components of the initial noise using the inherent power of the diffusion model. We refer to this method as ***FreeInit***, which substantially improves the generation quality without additional training or fine-tuning. The pipeline is illustrated in Fig. 7.

Denoise and Diffuse. During the inference process, an independent Gaussian noise ϵ is first initialized, which then undergoes the DDIM sampling process to yield a primary denoised latent z_0. Subsequently, we obtain the noisy latent z_T of the generated latent z_0 through the DDPM forward diffusion process, *i.e.*, adding noise to diffuse z_0 to z_T. Since z_T still preserves structural information from the denoised z_0 due to the information leakage, its low-frequency components have a better spatio-temporal correlation compared to ϵ. It is worth noting that, during this forward diffusion process, we have observed adding randomly sampled Gaussian noise could introduce significant uncertainty in the mid-frequency band, compromising the spatio-temporal correlation. Consequently, we opt to utilize the same original noise ϵ used in DDIM sampling when diffusing z_0 to z_T. The mathematical representation of this process is as follows:

$$\begin{aligned} z_T &= \sqrt{\overline{\alpha}_T} z_0 + \sqrt{1-\overline{\alpha}_T}\epsilon \\ &= \sqrt{\overline{\alpha}_T}(DDIM_{sample}(\epsilon)) + \sqrt{1-\overline{\alpha}_T}\epsilon, \end{aligned} \tag{8}$$

where $\overline{\alpha}_T$ is aligned with the β schedule used at training, *e.g.*, Stable Diffusion schedule.

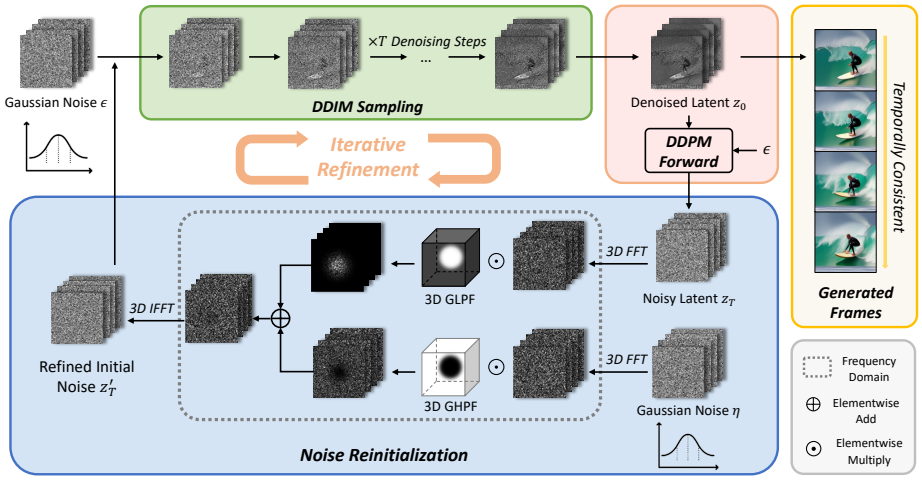


Fig. 7: Framework of FreeInit. FreeInit refines the low-frequency components of the initial noise in an iterative manner. During inference, a Gaussian Noise is first initialized and goes through the standard DDIM sampling process. The resulting denoised latent z_0 is then diffused using the original Gaussian Noise ϵ, through DDPM forward process. With the obtained noisy latent z_T which contains richer low-frequency information, a noise reinitialization process is further performed: z_T is firstly transformed into frequency domain through 3D FFT, then its spatio-temporal low-frequency components are fused with the high-frequency from a randomly sampled Gaussian noise η, bringing flexibility for refinement in higher frequency band. After transforming back to time domain, the refined z'_T is used as the initial noise for the next iteration.

Noise Reinitialization. To maintain alignment with the SNR distribution at training stage, we propose a noise reinitialization strategy, which is essential for the improvement of temporal consistency. Specifically, we employ a spatio-temporal frequency filter to combine the low-frequency components of the noise latent z_T with the high-frequency components of a random Gaussian noise η, resulting in a reinitialized noisy latent z'_T. This approach allows us to preserve essential information contained in z_T while introducing sufficient randomness in high-frequency to enhance visual details, complementing its improved low-frequency components. The mathematical operations are performed as follows:

$$\mathcal{F}^L_{z_T} = \mathcal{FFT}_{3D}(z_T) \odot \mathcal{H}, \tag{9}$$

$$\mathcal{F}^H_{\eta} = \mathcal{FFT}_{3D}(\eta) \odot (1 - \mathcal{H}), \tag{10}$$

$$z'_T = \mathcal{IFFT}_{3D}(\mathcal{F}^L_{z_T} + \mathcal{F}^H_{\eta}), \tag{11}$$

where $\mathcal{FFT}_{3D}$ is the Fast Fourier Transformation operated on both spatial and temporal dimensions, $\mathcal{H}$ is a spatial-temporal Low Pass Filter (LPF), $\mathcal{IFFT}_{3D}$ is the Inverse Fast Fourier Transformation.

Finally, this reinitialized noise z'_T serves as the starting point for a new round of DDIM sampling, facilitating the generation of frames with enhanced temporal consistency and visual appearance.

Iterative Refinement of Initial Noise. It is important to note that the aforementioned operations can be iteratively applied. At each iteration, the latent code undergoes improvements in spatio-temporal consistency by refining and preserving the low-frequency information from denoising. After that, it gains flexibility in the high-frequency domain through reinitialization, resulting in an improved initial noise for the subsequent iteration. In this iterative manner, the quality of the initial noise is progressively refined, effectively bridging the distribution gap between training and inference. Ultimately, this iterative process contributes to the overall enhancement of generation quality.

5 Experiments

5.1 Implementation Details

To evaluate the effectiveness and generalization of our proposed method, we apply the FreeInit strategy to three publically available diffusion based text-to-video models: AnimateDiff [10], ModelScope [45] and VideoCrafter [4]. Following [7, 36], we evaluate the inference performance with prompts from UCF-101 [41] and MSR-VTT [49] dataset. For UCF-101, we use the same prompt list as proposed in [7]. For MSR-VTT, we randomly sample 100 prompts from the test set for evaluation. We also incorporate diverse prompts from [17] for qualitative evaluations.

During inference, the parameters of frequency filter for each model are kept the same for fair comparison. Specifically, we use a Gaussian Low Pass Filter (GLPF) $\mathcal{H}_{\mathcal{G}}$ with a normalized spatio-temporal stop frequency of $D_0 = 0.25$. For each prompt, we first adopt the default inference settings of each model for a single inference pass, then apply 4 extra FreeInit iterations and evaluate the progress of generation quality. All FreeInit metrics in Quantitative Comparisons are computed at the 4^{th} iteration.

5.2 Evaluation Metrics

Temporal Consistency. To measure the temporal consistency of the generated video, we compute frame-wise similarity between the first frame and all succeeding $N-1$ frames. Noteworthily, one typical failure case in current video diffusion models is semantically close but visually inconsistent generation result. For example in Fig. 6 (a), all frames are semantically aligned ("biking"), but the appearance of the subject and background exhibits unsatisfactory consistency. Consequently, semantic-based features like CLIP [30] are not appropriate for evaluating the visual temporal consistency in video generation. Following previous studies [34], we utilize ViT-S/16 DINO [3, 27] to measure the visual similarities, denoted as the **DINO** metric. The metric is averaged on all frames.

Table 1: Quantitative Comparisons on Temporal Consistency. FreeInit significantly improves the temporal consistency of all baseline methods.

Method	DINO ↑	
	UCF-101	MSR-VTT
AnimateDiff [10]	85.24	83.24
AnimateDiff+FreeInit	**92.01**	**91.86**
ModelScope [45]	88.16	88.95
ModelScope+FreeInit	**91.11**	**93.28**
VideoCrafter [4]	85.62	84.68
VideoCrafter+FreeInit	**89.27**	**88.72**

Table 2: Quantitative Comparisons on Motion Quality. FreeInit also achieves the best motion quality metrics in most cases.

Method	FVD↓	MS($\lvert\Delta_{UCF}\rvert$ ↓)	DD($\lvert\Delta_{UCF}\rvert$ ↓)
AnimateDiff [10]	1340.96	89.31 (7.33)	97.03 (20.2)
AnimateDiff+FreeInit	**1032.47**	**96.60 (0.04)**	**75.30 (1.53)**
ModelScope [45]	785.30	95.00 (1.64)	**80.54 (3.71)**
ModelScope+FreeInit	**702.15**	**96.29 (0.35)**	68.61 (8.22)
VideoCrafter [4]	730.04	90.50 (6.14)	92.62 (15.79)
VideoCrafter+FreeInit	**675.39**	**93.45 (3.19)**	**83.27 (6.44)**

Motion Quality. To compensate the possible bias of the temporal consistency metric toward over-smoothed videos, we further provide metrics to evaluate the motion quality of the generated videos: **1) Fréchet Video Distance (FVD).** We follow [7] to perform zero-shot text-to-video generation on UCF-101 and sample 2,048 videos to compute the FVD between the generated distribution and real distribution. Smaller FVD means the distribution is closer to real videos. **2) Motion Smoothness (MS) and Dynamic Degree (DD).** Metrics from VBench [17] are utilized for further evaluation. We use the generated samples on UCF-101 prompts to compute the scores. The scores of real UCF videos (MS=96.64, DD=76.83) are set as a reference to compute the absolute difference $|\Delta_{UCF}|$. Smaller $|\Delta_{UCF}|$ means the motion quality is more similar to real videos.

5.3 Quantitative Comparisons

The quantitative comparison results are reported in Tabs. 1 and 2. According to Tab. 1, FreeInit significantly improves the temporal consistency of all base models on both prompt sets, by a large margin from 2.92 to 8.62. As for motion quality (shown in Tab. 2), the FVD metrics of all methods are also remarkably improved by FreeInit, indicating a general enhancement in realism. All MS scores are improved and become closer to realistic videos. Although the dynamic degree is decreased, their differences with the ground truth UCF videos mostly become smaller (*e.g.*, AnimateDiff from 20.2 to 1.53). This proves the generated videos are not over-smoothed, but instead become closer to real video distributions.

We also conduct a User Study to evaluate the results through Temporal Consistency, Text Alignment and Overall Quality, which can be refer to the Supplementary File.

Fig. 8: Qualitative Comparisons. We apply FreeInit to different base models and inference with diverse text prompts. FreeInit significantly improves the temporal consistency and the subject appearance of the generated videos.

5.4 Qualitative Comparisons

Qualitative comparisons are shown in Fig. 8. Our proposed FreeInit significantly improves the temporal consistency as well as visual quality. For example, with text prompt "a musician playing the flute", performing FreeInit effectively fix the temporally unstable artifacts exhibited in vanilla AnimateDiff. More qualitative results are listed in the Supplementary File.

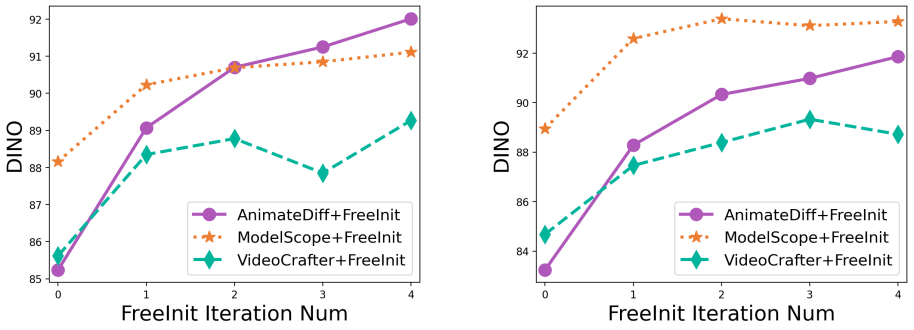
5.5 Ablation Study

In this section, we quantitatively evaluate the design choices and parameters of FreeInit. Qualitative results can be referred to the Supplementary File.

Influence of Noise Reinitialization and Filter Selection. To evaluate the importance of Noise Reinitialization in the frequency domain and the choice of filter, we run two FreeInit variants on both datasets with all three base models. Firstly, Noise Reinitialization is totally skipped, *i.e.*, the noisy latent z_T after DDPM Forward Pass is directly used as initial noise for sampling. Secondly, the frequency filter used for Noise Reinitialization is changed from GLPF to ILPF, with the same stop frequency 0.25. The metrics in Tab. 3 clearly demonstrate that Noise Reinitialization is crucial for improving temporal consistency. Also, replacing the soft Gaussian filter GLPF with the hard Ideal filter ILPF leads to a performance drop, which reveals the importance of also introducing moderate randomness into mid-frequency and low-frequency components. More detailed discussions are in the Supplementary File.

Table 3: Ablation Study on Noise Reinitialization (NR). Removing NR or changing Gaussian Low Pass Filter (GLPF) to Ideal Low Pass Filter (ILPF) leads to non-optimal results. *ModelName** refers to *Model+FreeInit*.

Method	DINO ↑	
	UCF-101	MSR-VTT
AnimateDiff* w/o NR	86.77	85.18
AnimateDiff* w/ NR-ILPF	87.53	86.17
AnimateDiff* w/ NR-GLPF	**92.01**	**91.86**
ModelScope* w/o NR	88.20	90.90
ModelScope* w/ NR-ILPF	89.04	90.93
ModelScope* w/ NR-GLPF	**91.11**	**93.28**
VideoCrafter* w/o NR	86.09	87.11
VideoCrafter* w/ NR-ILPF	87.53	88.01
VideoCrafter* w/ NR-GLPF	**89.27**	**89.33**



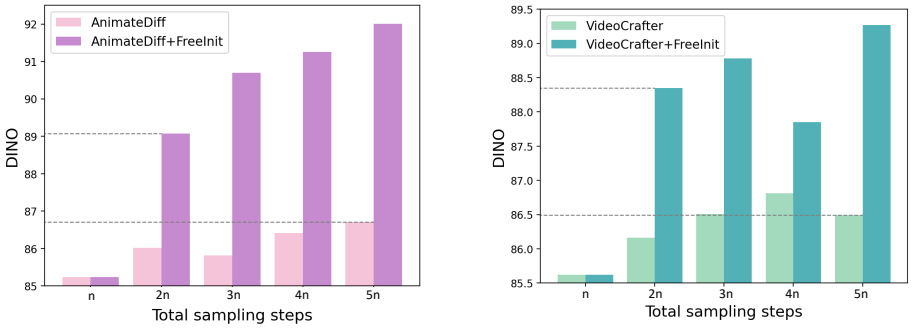
(a) UCF-101 **(b)** MSR-VTT

Fig. 9: Ablation Study on Iteration Number. We report the DINO scores under different FreeInit iteration numbers on (a) UCF-101 and (b) MSR-VTT. More iteration steps mostly leads to better temporal consistency, and the most significant improvement is observed at the 1^{st} iteration.

Influence of Iteration Steps. We show the influence of FreeInit iteration step number in Fig. 9. It can be observed that the temporal consistency consistently increases with the iteration step, thanks to the gradually refined initial noise. Notably, the largest temporal consistency improvement for each model comes from the *1st* iteration, where FreeInit is applied for the first time. This is because at the *0-th* iteration, the initial noise is non-correlated Gaussian noise, while at the *1st* iteration, low-frequency information is injected into the noise for the first time, largely eliminating the gap between inference noise and training noise.

5.6 Further Discussion

Comparison with Same Inference Step without FreeInit. Since FreeInit uses more than one DDIM sampling pass, it is natural to ask if the quality improvement is due to the increased sampling steps. To answer this question, we compare FreeInit with the typical DDIM sampling strategy using the same total inference steps. As shown in Fig. 10, trivially increasing the DDIM sampling



(a) AnimateDiff (n=25) **(b)** VideoCrafter (n=50)

Fig. 10: Comparison with Same Sampling Steps without FreeInit. We analyze if increasing the DDIM sampling steps for baseline methods would help to improve the temporal consistency on UCF-101. For all base models, the vanilla inference with $5n$ steps is inferior to incorporating FreeInit with $2n$ steps. This indicates that FreeInit is not equivalent to trivially increasing the DDIM sampling steps.

steps only brings little improvement in temporal consistency. Notably, with just one extra FreeInit iteration (total $2n$ steps), the temporal consistency becomes even better than using $5n$ vanilla DDIM sampling steps that require $\times 2.5$ time cost. This further proves the importance of refining initial noise at inference time: *a good beginning matters more than struggling with a bad initial state.*

Limitations. As an iterative method, a natural drawback of FreeInit is the increased sampling time. However, incorporating FreeInit leads to much higher performance gain compared to spending more time using the common sampling strategy (Fig. 10). Furthermore, this issue can be mitigated through a coarse-to-fine sampling strategy. We explain more details and discuss more about the limiations and potential negative societal impacts in the Supplementary File.

Broader Applications. Since the training-inference initialization gap is a common issue, FreeInit is applicable to not only video diffusion models, but also other kinds of diffusion models, *e.g.*, text-to-image models like SDXL [28]. Results and discussions are provided in the Supplementary File.

6 Conclusion

In this paper, we identify an implicit training-inference gap in the noise initialization of video diffusion models that causes degenerated inference quality: 1) the frequency distribution of the initial noise's SNR is different between training and inference; 2) the denoising process is significantly affected by the low-frequency components of initial noise. Based on these observations, we propose FreeInit, which improves temporal consistency through the iterative refinement of the spatial-temporal low-frequency component of the initial noise during inference. This narrows the initialization gap between training and inference. Extensive quantitative and qualitative experiments on various text-to-video models and text prompts demonstrate the effectiveness of our proposed FreeInit.

Acknowledgement

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
2. Brooks, T., Hellsten, J., Aittala, M., Wang, T.C., Aila, T., Lehtinen, J., Liu, M.Y., Efros, A., Karras, T.: Generating long videos of dynamic scenes. In: NeurIPS (2022)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
4. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
5. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS (2021)
6. Everaert, M.N., Fitsios, A., Bocchio, M., Arpa, S., Süsstrunk, S., Achanta, R.: Exploiting the signal-leak bias in diffusion models. arXiv preprint arXiv:2309.15842 (2023)
7. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: ICCV (2023)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
9. Gu, J., Wang, S., Zhao, H., Lu, T., Zhang, X., Wu, Z., Xu, S., Zhang, W., Jiang, Y.G., Xu, H.: Reuse and diffuse: Iterative denoising for text-to-video generation. arXiv preprint arXiv:2309.03549 (2023)
10. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
11. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible diffusion modeling of long videos. arXiv preprint arXiv:2205.11495 (2022)
12. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
13. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
16. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: CogVideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)

17. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. arXiv preprint arXiv:2311.17982 (2023)
18. Jiang, Y., Yang, S., Koh, T.L., Wu, W., Loy, C.C., Liu, Z.: Text2Performer: Text-driven human video generation. In: ICCV (2023)
19. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS (2021)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR (2020)
22. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
23. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. arXiv preprint arXiv:2305.08891 (2023)
24. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: cvpr (2022)
25. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: VideoFusion: Decomposed diffusion models for high-quality video generation. In: CVPR (2023)
26. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
29. Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., Liu, Z.: Freenoise: Tuning-free longer video diffusion via noise rescheduling. arXiv preprint arXiv:2310.15169 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
33. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)

36. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
37. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: CVPR (2022)
38. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
40. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
41. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
42. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. arXiv preprint arXiv:2104.15069 (2021)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
44. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
45. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
46. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
47. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions (2021)
48. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation (2021)
49. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016)
50. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation (2023)
51. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
52. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models (2023)