FreeInit: Bridging Initialization Gap in Video Diffusion Models
–*Supplementary File*–

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

In this ***Supplementary File***, we first explain the SNR distribution computation in Sec. A, then tabulate the user study results in Sec. B. More detailed discussions on Noise Reinitialization, filter parameters selection and iteration steps are provided in Sec. C. Broader applications (*e.g.* SDXL) are discussed in Sec. D. We list more implementation details in Sec. E. More qualitative comparisons are illustrated in Sec. F to visualize the performance of FreeInit. We further discuss some limitations of FreeInit and its possible social impact in Sec. G and Sec. H. Source code, demo video and more visual comparisons can be found in our project page: https://tianxingwu.github.io/pages/FreeInit/

A Signal-to-Noise Ratio Distribution

In this section, we explain how we derive the frequency SNR distribution of z_t in manuscript Sec. 3.2.

Mathematically, SNR is defined as the ratio of the power of the signal P_{signal} to the power of the noise P_{noise}:

$$SNR = \frac{P_{signal}}{P_{noise}} = \frac{A^2_{signal}}{A^2_{noise}} \tag{1}$$

Where A denotes the amplitude of the signal and the noise.

To measure the frequency distribution of the hidden information in the noisy latent z_t during training, we apply 3D Fourier Transformation $\mathcal{FFT}_{3D}$ to both the clean latent z_0 and the Gaussian noise ϵ:

$$\mathcal{F}_{z_0} = \mathcal{FFT}_{3D}(z_0), \mathcal{F}_{\epsilon} = \mathcal{FFT}_{3D}(\epsilon) \tag{2}$$

The amplitude spectrum can then be derived with the absolute value of the frequency-domain representation:

$$\mathcal{A}_{z_0} = |\mathcal{F}_{z_0}|, \mathcal{A}_{\epsilon} = |\mathcal{F}_{\epsilon}| \tag{3}$$

According to Eqn. 4 in manuscript and the linear property of the Fourier transform, the full-band SNR of z_t is derived as:

$$SNR(z_t) = \frac{(\sqrt{\hat{\alpha}_t} A_{z_0})^2}{(\sqrt{1-\hat{\alpha}_t} A_{\epsilon})^2} = \frac{\hat{\alpha}_t}{1-\hat{\alpha}_t} \frac{A^2_{z_0}}{A^2_{\epsilon}} \tag{4}$$

Consider a frequency band Φ with the spatio-temporal frequency range in $\{(f_s^L, f_s^H), (f_t^L, f_t^H)\}$, the SNR of z_t in this frequency band can be calculated using a band-pass filter (BPF), or approximate by summing the amplitudes in the corresponding spatio-temporal range. Converting to logarithm scale, the SNR for frequency band Φ is finally derived as:

$$SNR_{dB}(z_t, \Phi) = 10 \log_{10} \frac{\hat{\alpha}_t}{1-\hat{\alpha}_t} \frac{\sum_{f_s^L}^{f_s^H} \sum_{f_t^L}^{f_t^H} A_{z_0}^2}{\sum_{f_s^L}^{f_s^H} \sum_{f_t^L}^{f_t^H} A_\epsilon^2} \tag{5}$$

B User Study

We conduct a User Study to further evaluate the influence of FreeInit. We randomly select 72 diverse text prompts for the test models (VideoCrafter [1], ModelScope [8] and AnimateDiff [4]), and ask 42 participants to vote for the generation results. Specifically, each participant is provided with the text prompt and a pair of synthesized videos, one generated from the vanilla model and the other one with FreeInit. Then the participants vote for the video that they consider superior for Temporal Consistency, Text-Video Alignment and Overall Quality, respectively. The average vote rates are shown in Tab. A1. The majority of the votes go to the category using FreeInit under all evaluation metrics, which indicates that FreeInit consistently improves the quality of video generation.

Table A1: User Study. Each participant votes for the image that they consider superior for Temporal Consistency, Text-Video Alignment and Overall Quality, respectively.

Method	Temporal Consistency	Text-Video Alignment	Overall Quality
VideoCrafter [1]	13.10%	20.24%	18.45%
VideoCrafter+FreeInit	**86.90%**	**79.76%**	**81.55%**
ModelScope [8]	19.05%	23.81%	18.45%
ModelScope+FreeInit	**80.95%**	**76.19%**	**81.55%**
AnimateDiff [4]	9.33%	24.30%	15.97%
AnimateDiff+FreeInit	**90.67%**	**75.70%**	**84.03%**

C More Discussions on Ablation Study

Influence of Noise Reinitialization and Filter Selection. The qualitative results generated by the vanilla AnimateDiff model and three ablation variants of FreeInit are illustrated in Fig. A1. Aligned with the conclusion in the quantitative results in the manuscript, the visual results indicate that the Noise Reinitialization process is crucial for the refinement of temporal consistency. As depicted in

(a) w/o FreeInit (b) FreeInit w/o NR

(c) FreeInit w/ NR-ILPF (d) FreeInit w/ NR-GLPF **(ours)**

Fig. A1: Ablation Study on Noise Reinitialization. Using Noise Reinitialization with a proper frequency filter is crucial for generating temporally consistent results.

Table A2: Ablation on Cut-off Frequency. Our setting achieves a balanced dynamics-consistency trade-off.

D_0	**DINO↑**	**MS($\lvert\Delta_{UCF}\rvert\downarrow$)**	**DD($\lvert\Delta_{UCF}\rvert\downarrow$)**
0.125	91.54	92.83 (3.81)	89.52 (12.69)
0.25 (ours)	92.07	**96.80 (0.16)**	**75.24 (1.59)**
0.5	**93.91**	98.70 (2.06)	39.05 (37.78)

Fig. A1, the color of the boxers' clothes and the background suffer from large inconsistencies in the original generation results. Adding the FreeInit refinement loop without Noise Reinitialization still leads to unsatisfactory appearance and undesirable temporal consistency, as no randomness is provided for developing better visual details that complement well with the refined low-frequency components. With Noise Reinitialization performed by Ideal Low Pass Filter (ILPF), the consistency of the subject appearance gains clear improvement, but the large noises (*e.g.*, the yellow area) are still not removed. In comparison, Reinitialization with Gaussian Low Pass Filter (GLPF) is able to remove the large noises, as it introduces an abundant amount of randomness into both mid and high-frequency, creating room for fixing large visual discrepancies. Despite gaining results with better temporal consistency, we also find using Gaussian Low Pass Filter sometimes leads to over-smoothed frames. To alleviate this side effect, the Butterworth Low Pass Filter can be utilized to replace the GLPF for keeping a balance between temporal consistency and visual quality.

Choice of Cut-off Frequency. The ablation study on the cut-off frequency of GLPF is shown in Tab. A2. We test different values of D_0 with AnimateDiff+FreeInit setting. When D_0 increases, the low-frequency component becomes more dominant thus leads to an increase in temporal consistency and motion smoothness. However, large D_0 also harms the dynamic degree. We find $D_0 = 0.25$ achieves an optimal elbow point with the dynamics-consistency trade-off, achieving the most balanced quality. Users may adjust the parameter according to the property of specific base model and their special needs.

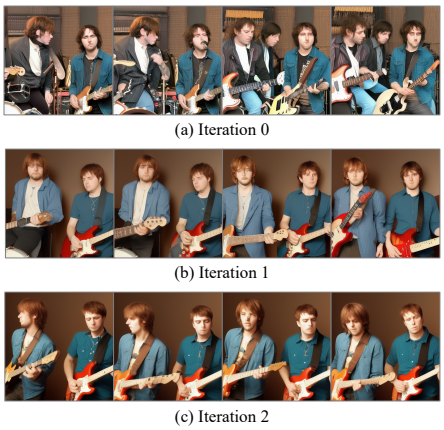


Fig. A2: Ablation Study on Iteration Steps. The appearance of the players and instruments becomes more consistent after each iteration. However, it is worth mentioning that the largest consistency leap takes place in the first iteration.

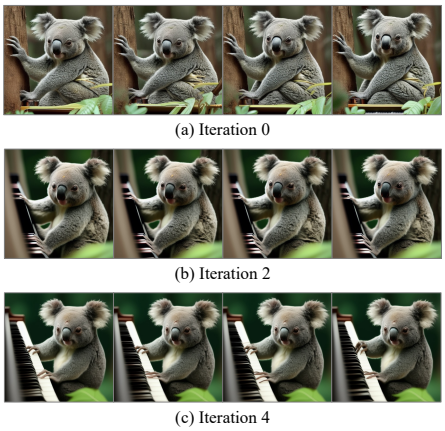


Fig. A3: Semantic Growth. The frames are generated with the text prompt "A koala bear is playing piano in the forest". The missing semantics "playing piano" gradually grows with the FreeInit iteration.

Influence of Iteration Steps. Since the low-frequency components of the latent are refined during each FreeInit iteration, more iteration steps normally lead to generation results with better temporal consistency. Nonetheless, the most significant improvement comes from the 1st refinement iteration (iteration 1), as shown in Fig. A2. We also observe that the missing semantics (*e.g.*, "playing piano", as depicted in Fig. A3) can gradually grow with the FreeInit iteration, thanks to the refined low-frequency components and the randomness introduced by Noise Reinitialization.

D Broader Applications

Since our discovered 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.

To enable FreeInit on image models, we remove the temporal dimension of the original spatio-temporal frequency decomposition operation in FreeInit, and use a 2D GLPF with $D_0 = 0.125$ to implement the Noise Reinitialization.

We show visual results of adding FreeInit to SDXL [7] in Fig. A4. By iteratively refining the initial noise, FreeInit helps SDXL to generate very dark and bright images with improved text alignment and high visual quality. This functionality is similar to the approach proposed in [6], while we require no additional training or fine-tuning to narrow the SNR gap.

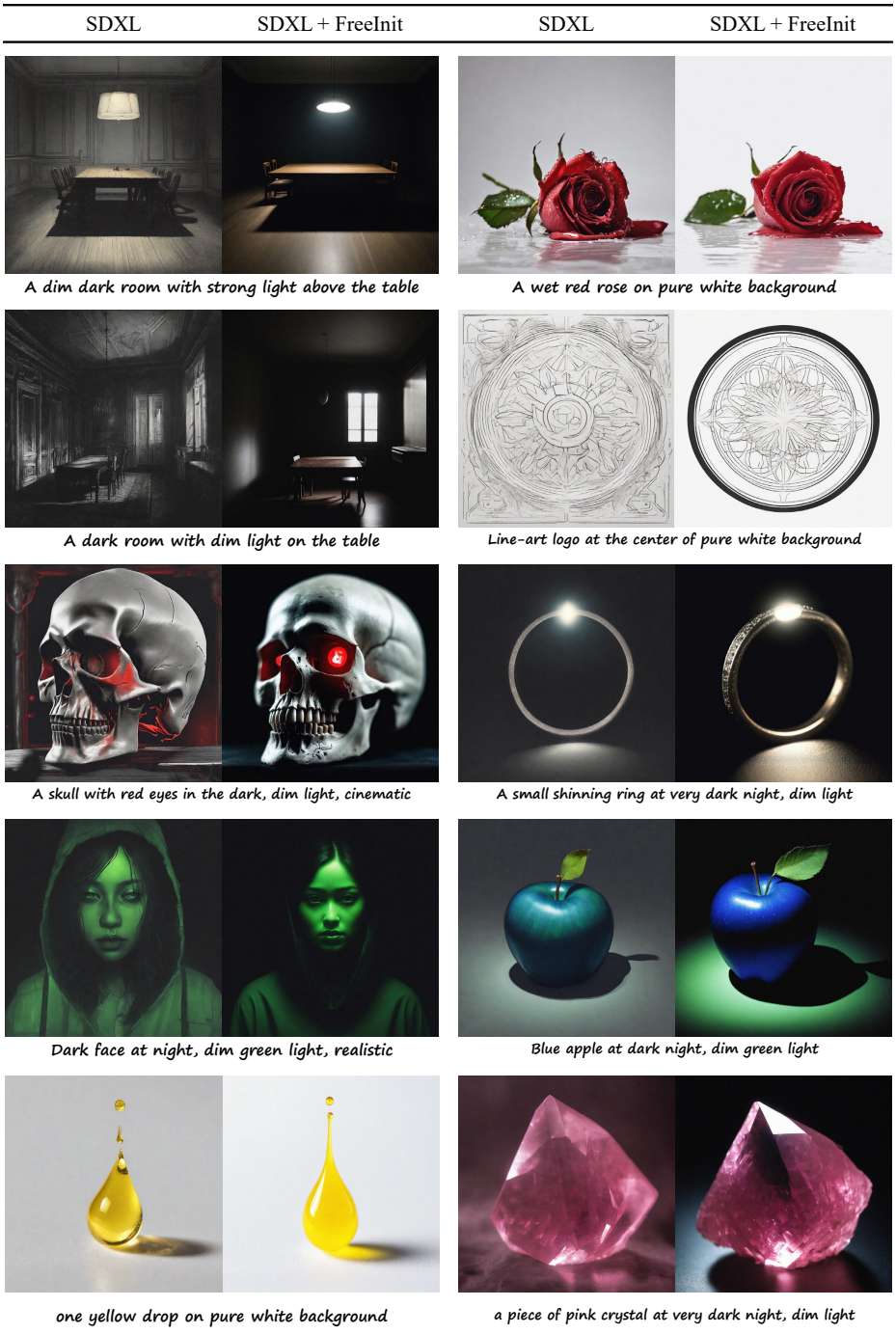


Fig. A4: FreeInit for Text-to-Image Diffusion Models. It enables SDXL [7] to generate very dark/bright images with better text alignment and high visual quality.

E Implementation Details

Base Models. Three open-sourced text-to-video models are used as the base models for FreeInit evaluation. For VideoCrafter [1], the VideoCrafter-v0.9 Base T2V model based on the latent video diffusion models (LVDM) [5] is adopted. For ModelScope [8], we utilize the *diffusers* [3] implementation of the text-to-video pipeline, with the text-to-video-ms-1.7b model. For AnimateDiff [4], we use the mm-sd-v14 motion module with the Realistic Vision V5.1 LoRA model for evaluation.

Inference Details. Experiments on VideoCrafter and ModelScope are conducted on 256×256 spatial scale and 16 frames, while experiments on AnimateDiff are conducted on a video size of 512×512, 16 frames. During the inference process, we use classifier-free guidance for all experiments including the comparisons and ablation studies, with a constant guidance weight 7.5. All experiments are conducted on a single Nvidia A100 GPU.

F More Qualitative Comparisons

Extra qualitative results on AnimateDiff [4], ModelScope [8] and VideoCrafter [1] are provided in Fig. A8- A13.

We also include visual comparisons on the more recent video diffusion model VideoCrafter2 [2] in Fig. A14. As can be observed in the results, although VideoCrafter2 generally archives better generation quality, temporal inconsistencies still exist, and FreeInit is able to mitigate the issue.

For more visual results in video format, please refer to our project page.

G Limitations

Inference Time. Since FreeInit is an iterative method, a natural drawback is the increased sampling time. To avoid unnecessary time cost, we only use around 2-3 extra FreeInit iterations in practice, since the largest performance improvement mainly comes from the 1st refinement iteration (Fig. A2, Manuscript Fig.9), and that the performance increase saturates after 2-3 iterations (Manuscript Fig.9). Along with numerical optimization, we can sample a high-resolution $512 \times 512 \times 16$ video with the optimized AnimateDiff+FreeInit in less than 30s on a single A100 GPU.

This issue can be further mitigated through a **Coarse-to-Fine Sampling** Strategy. Specifically, the DDIM sampling steps of each FreeInit iteration can be reduced according to the iteration step: early FreeInit iterations use fewer DDIM sampling steps for a coarse refinement of the low-frequency components, while the latter iterations perform more steps for detailed refinement. A simple linear scaling strategy can be used for Coarse-to-Fine Sampling:

$$T_i = [\frac{T}{N}(i+1)] \tag{6}$$

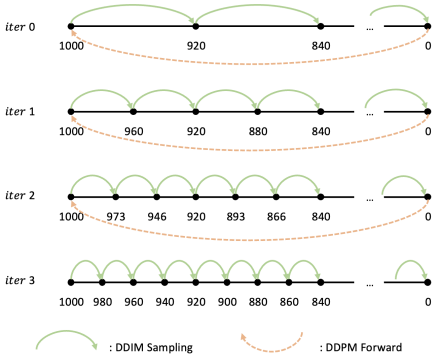


Fig. A5: Coarse-to-Fine Sampling for Fast Inference. Early FreeInit iterations use fewer DDIM sampling steps, while latter iterations perform more steps for detailed refinement.

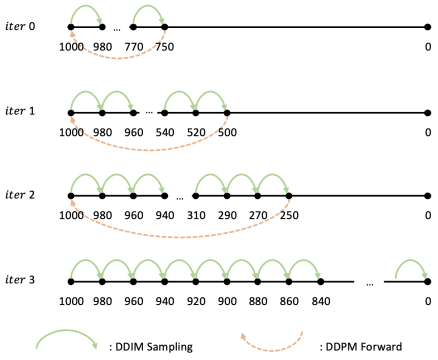


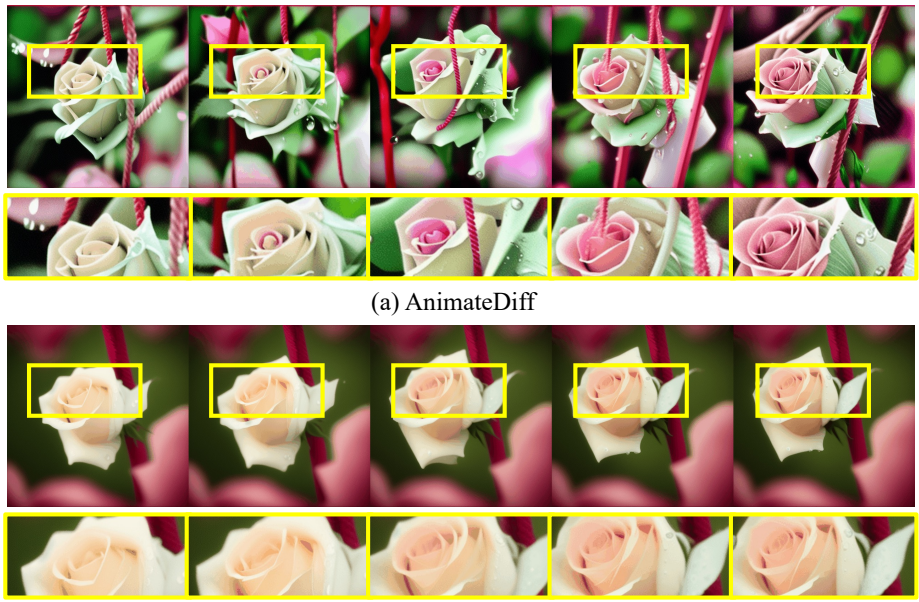
Fig. A6: Another Design Choice for the Coarse-to-Fine Sampling. We empirically find that this strategy is sub-optimal compared to the other one in several cases.

Where T_i is the DDIM steps for iteration i, T is the commonly used DDIM steps (*e.g.*, 50), and N is the number of FreeInit iterations. For instance, the fast Sampling process for $T = 50$, $N = 4$ is illustrated in Fig. A5. Besides, we also test another fast sampling strategy, as illustrated in Fig. A6. The key idea is to keep the step size fixed but apply an early-stop at early iterations. However, we find this strategy is sub-optimal compared to the former one.

Failure Cases. In some cases where the video includes small and fast-moving foreground objects, performing FreeInit occasionally leads to the distinction of the object. As shown in the example Fig. A7, although the temporal consistency of the rose is enhanced, the small object "waterdrop" almost vanishes. This is because the iterative low-frequency refinement strategy tends to guide the generation towards the more stable low-frequency subjects. As the iteration progresses, the enhanced semantics within the initial noise's lower frequencies increasingly take more control of the generation process, causing a partial loss of control from the text condition. Regarding this, the users can choose to use fewer FreeInit iterations, tune the frequency filter parameters, or adjust the classifier-free guidance weight to balance the trade-off between depicting small objects and enhancing temporal consistency, according to their specific needs and preferences.

H Potential Negative Societal Impacts

FreeInit is a research focused on improving the inference quality of existing video diffusion models without favoring specific content categories. Nonetheless, its application in aiding other video generation models could potentially be exploited for malicious purposes, resulting in the creation of fake content.



(a) AnimateDiff

(b) AnimateDiff + FreeInit

Fig. A7: Failure Case. With input prompt "a rose swing in the wind with waterdrops", performing FreeInit improves temporal consistency but falsely removes the fast, small foreground object (waterdrops).

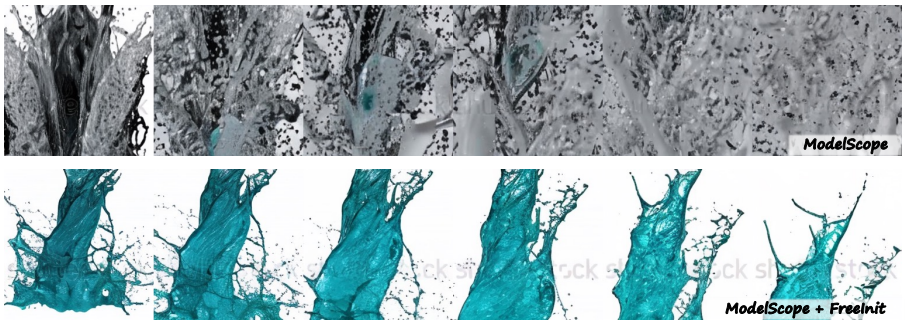
A corgi is playing drum kit.

Vampire makeup face of beautiful girl, red contact lenses.

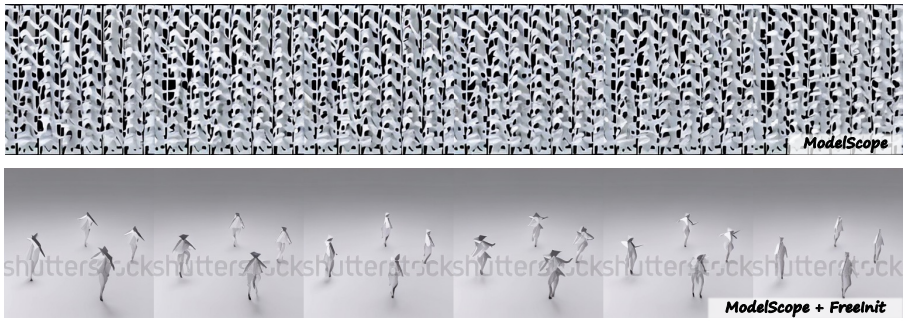
A cat wearing sunglasses and working as a lifeguard at a pool.

Fig. A8: More Qualitative Results of FreeInit on AnimateDiff.

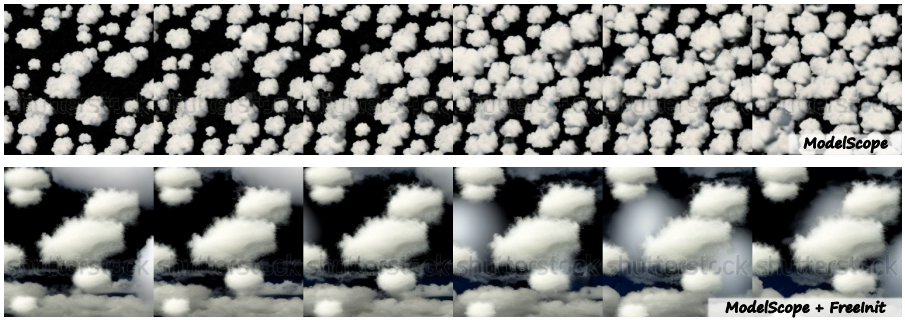
Fig. A9: More Qualitative Results of FreeInit on AnimateDiff.



Splash of turquoise water in extreme slow motion, alpha channel included.



Origami dancers in white paper, 3D render, on white background, studio shot, dancing modern dance.



An animated painting of fluffy white clouds moving in sky.

Fig. A10: More Qualitative Results of FreeInit on ModelScope.

A beautiful coastal beach in spring, waves lapping on sand by Vincent van Gogh.

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

A bigfoot walking in the snowstorm.

Fig. A11: More Qualitative Results of FreeInit on ModelScope.

Turtle swimming in ocean.

Campfire at night in a snowy forest with starry sky in the background.

A cute raccoon playing guitar in a boat on the ocean.

Fig. A12: More Qualitative Results of FreeInit on VideoCrafter.

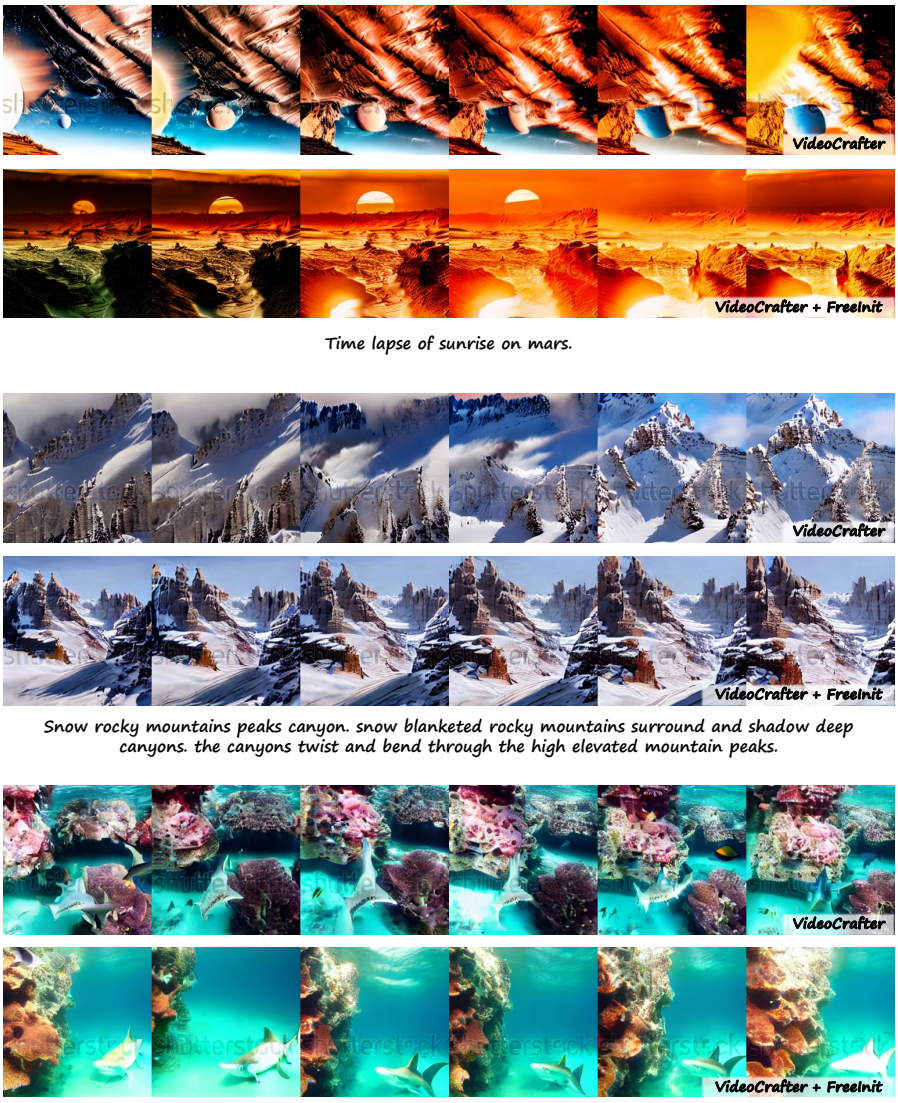


Fig. A13: More Qualitative Results of FreeInit on VideoCrafter.

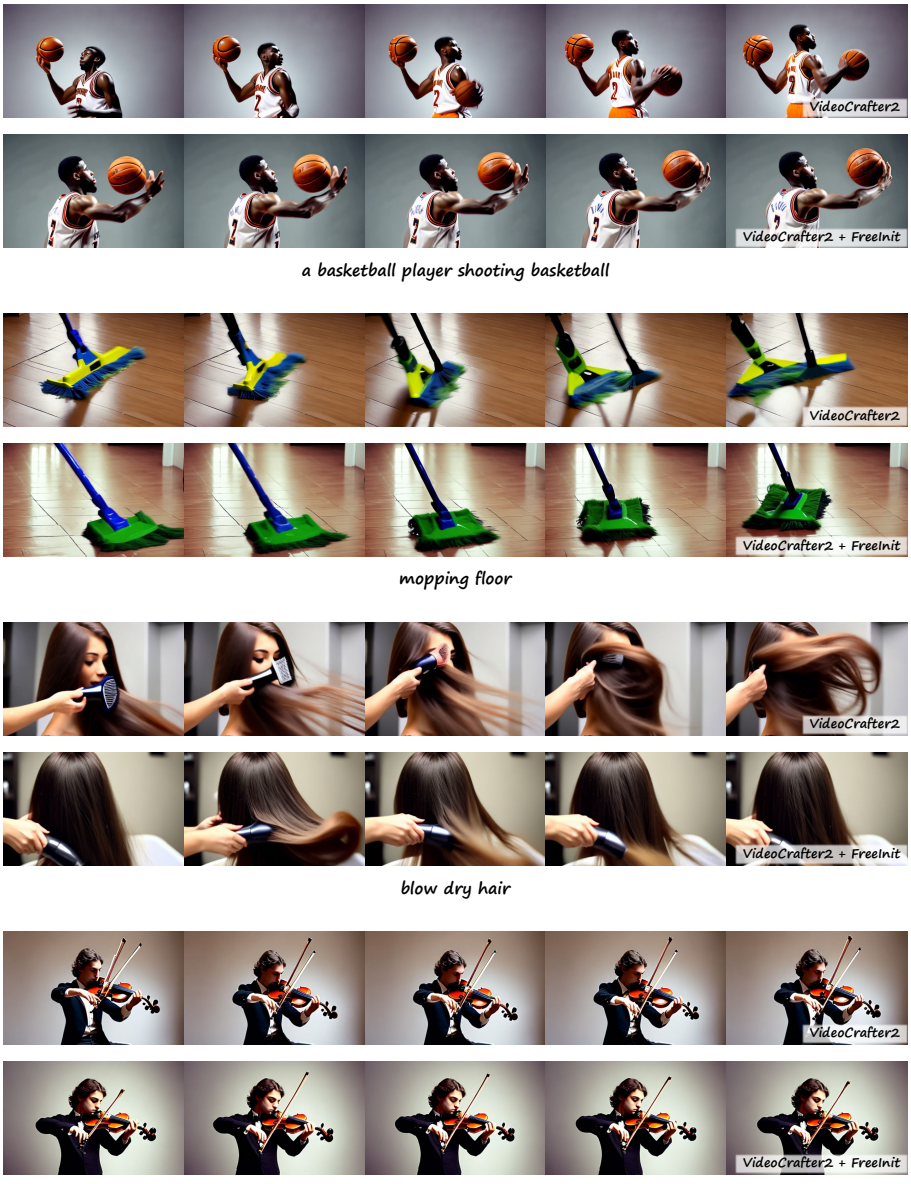


Fig. A14: More Qualitative Results of FreeInit on VideoCrafter2.

References

1. 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)
2. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR (2024)
3. Face, H.: Diffusers. https://huggingface.co/docs/diffusers/index
4. 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)
5. 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)
6. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. arXiv preprint arXiv:2305.08891 (2023)
7. 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)
8. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)