DreamMotion: Space-Time Self-Similar Score Distillation for Zero-Shot Video Editing

Supplementary Material

Hyeonho Jeong[1], Jinho Chang[1], Geon Yeong Park[2], and Jong Chul Ye[1,2]

[1] Kim Jaechul Graduate School of AI, KAIST, South Korea
[2] Dept. of Bio and Brain Engineering, KAIST, South Korea
{hyeonho.jeong, jinhojsk515, pky3436, jong.ye}@kaist.ac.kr
Project page: https://hyeonho99.github.io/dreammotion

A Related Work

A.1 Video Editing using Diffusion Models

Creating videos from textual descriptions necessitates ensuring realistic and temporally consistent motion, posing a unique set of challenges compared to text-driven image generative scenarios. Before the advent of publicly accessible text-to-video diffusion models, Tune-A-Video [30] was at the forefront of one-shot-based video editing. They proposed to inflate the image diffusion model to the pseudo video diffusion model by appending temporal modules to the image diffusion model [22] and reformulating spatial self-attention into spatio-temporal self-attention, facilitating inter-frame interactions. However, the inflation often falls short of achieving consistent and complete motion, as motion preservation relies implicitly on the attention mechanism during inference. Thus, the attention projection matrices within U-Net are often fine-tuned on the input videos [13,30,36]. Utilizing explicit visual signals to steer the video denoising process is another common technique. Pix2Video [1] and FateZero [21], for instance, inject intermediate attention maps during the editing phase, which are derived during the input video inversion. Others leverage pre-trained image adapter networks for structurally consistent video generation. A notable example is ControlNet [33], which has been modified to accommodate a series of explicit structural indicators such as depth and edge maps. Ground-A-Video [10] takes this a step further by adapting both ControlNet and GLIGEN [11] for video editing, utilizing spatially-continuous depth maps and spatially-discrete bounding boxes.

Despite the availability of open-source text-to-video (T2V) diffusion models [2, 24, 26, 27, 32], recent endeavors frequently adopt a self-supervised strategy of fine-tuning pre-trained video generative models on an input video, to accurately capture intricate, real-world motion. More specifically, several studies attempt to disentangle the appearance and motion elements of videos during the self-supervised fine-tuning. For example, [29, 35, 37] split the fine-tuning phase into two distinct pathways: one dedicated to integrating the subject's appearance

into spatial modules, and the other aimed at embedding motion dynamics of a video into temporal modules within the T2V model. Additionally, other studies [9,14,18] attempt to extract and learn motion information from a single or a few reference videos. VMC [9], for instance, proposes to distill the motion within a video by calculating the residual vectors between consecutive frames, and refine temporal attention layers in cascaded video diffusion models.

Distinct from aforementioned approaches, DreamMotion circumvents the conventional ancestral sampling and employs Score Distillation Sampling [20] for editing appearance elements within a video.

A.2 Visual Generation using Score Distillation Sampling

Score Distillation Sampling (SDS) [20], also known as Score Jacobian Chaining, has become the go-to method for text-to-3D generation in recent years [3, 8, 12, 15, 19, 23, 25, 28]. DreamFusion [20] first proposed to distill the generative prior of pre trained text-to-image models and optimize a parametric image synthesis model, such as NeRF [16]. Despite its success, SDS often produces images that are overly saturated, blurry, and lack detail, largely due to the use of high CFG values [28]. To address these challenges, a range of derivative methods have been proposed [7, 8, 12, 15, 17, 28]. Specifically, in the context of accurate image editing, DDS [7] incorporates an additional reference branch with corresponding text to refine the noisy gradient of SDS. Hifa [38], instead, utilizes an estimated clean image rather than the predicted noise to compute denoising scores. In our work, we employ a straightforward yet effective mask condition to refine DDS-generated gradients, allowing us to inject particular appearance into the video. We further ensures the preservation of the video's original structure and motion through the novel regularization of space-time self-similarity alignment.

B Technical Details

For the sampling of timestep t to derive $\boldsymbol{x}_t^{1:N}$ and $\hat{\boldsymbol{x}}_t^{1:N}$, we restrict t to the range $t \sim \mathcal{U}(0.05, 0.95)$, in line with DDS's official implementation[3]. For the extraction of attention key features from video diffusion U-Net, we specifically select the self-attention layers within its decoder part. In the non-cascaded video diffusion experiments, we utilize Zeroscope[4] [24], a diffusion model that operates in latent-space rather than pixel-space. Practically, this means the video frames are initially encoded into latent representations by VAEs, and then our proposed optimizations take place within this latent space. Conversely, in experiments involving the cascaded video diffusion framework, we select Show-1[5] [32], where the keyframe generation UNet of Show-1 uses a pixel-space diffusion. As a result, the video frames stay in pixel space, with optimizations carried out directly within this domain.

[3] https://github.com/google/prompt-to-prompt/blob/main/DDS_zeroshot.ipynb
[4] https://huggingface.co/cerspense/zeroscope_v2_576w
[5] https://huggingface.co/showlab/show-1-base

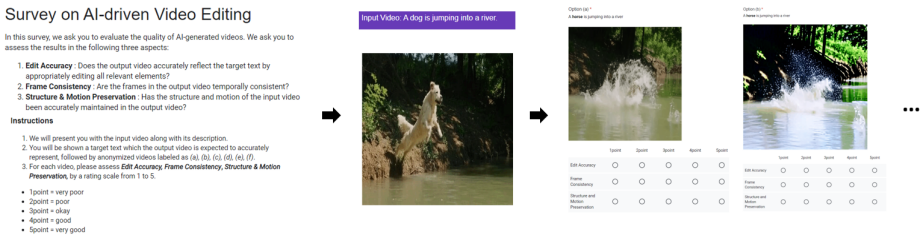

Fig. 1: Interface of human evaluation.

To produce output videos using Tune-A-Video [30], ControlVideo [34], Control-A-Video [4], and TokenFlow [6], we utilized the official github repositories along with their default hyperparameters. The results from Gen-1 [5] were generated using their web-based product. Given that Gen-1 generates videos with temporally extended sequences, including duplicated frames, we removed these repeated frames when calculating CLIP-based frame consistency to ensure a fair evaluation. However, for the human evaluation, the outputs from Gen-1 were used as is, without any modifications.

C User Study Interface

We carried out human evaluations to assess various methods based on three key aspects: Edit Accuracy, Frame Consistency, and Sturcture & Motion Preservation. Initially, we present the input video alongside its text description, as shown in Figure 1. Subsequently, we display target text with anonymized videos generated by each method and ask participants to evaluate them across the aforementioned three criteria. The human evaluation results, detailed in Table 1 of the manuscript, unequivocally highlight the superiority of DreamMotion in both video diffusion frameworks.

D Additional Comparison

We additionally compared our method with two video editing techniques specifically designed for localized editing: Video-P2P [13] and Diffusion-Motion-Transfer (DMT) [31]. For qualitative comparison, see Fig. 2. For quantitative comparison in Tab. 1, we employed tracking-based motion fidelity score [31] and framewise LPIPS [13] to evaluate spatial consistency.

	Video-P2P	DMT	Ours (Zeroscope)
Motion-Fidelity	0.7384	0.8697	**0.9259**
Frame-LPIPS	0.3395	0.3078	**0.3042**

Table 1: Additional quantitative comparison with DMT and Video-P2P.

Fig. 2: Additional qualitative comparison with DMT and Video-P2P.

E Additional Results

This section is dedicated to presenting additional outcomes of DreamMotion. Figure 3 offers a comprehensive view of the results from Figure 6 in the main paper, demonstrating the effect of masking DDS-driven gradients. Annotations within the input video frames indicate the masks used. In Figure 4, we present the progress of DreamMotion optimization by visualizing intermediate output videos. Figures 5, 6, and 7 showcase input and corresponding edited videos generated with DreamMotion on Zeroscope, using various target prompts. To accommodate space constraints, only odd or even frames from 16-frame videos are selected for display. Figures 8, 9, and 10 feature videos edited by DreamMotion on the Show-1 Cascade model [32], with the left columns displaying 8-frame input videos and the adjacent columns showing 29-frame output videos. Our qualitative results are uploaded on our project page.

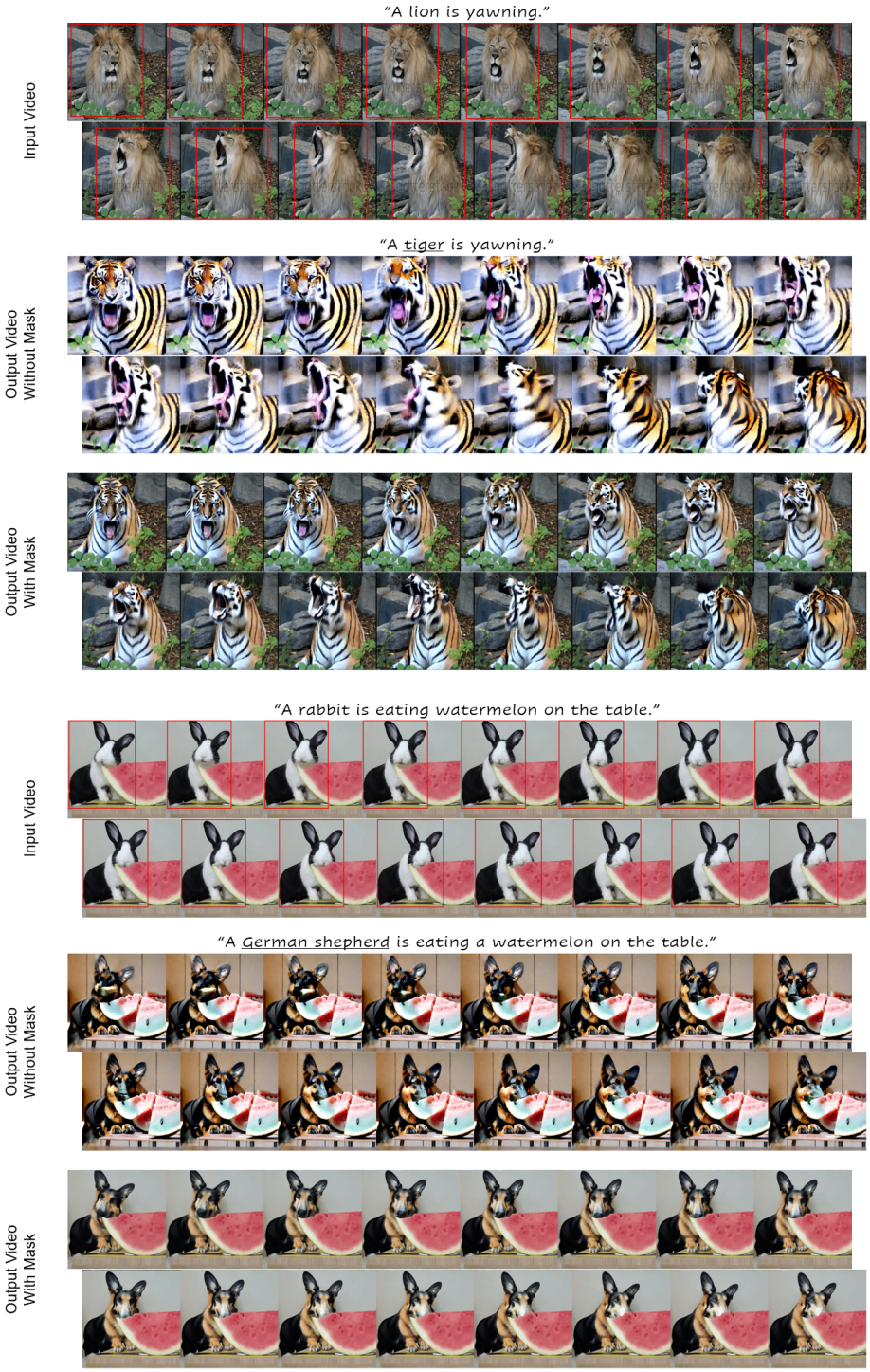

Fig. 3: Video optimization with and without masking gradients.

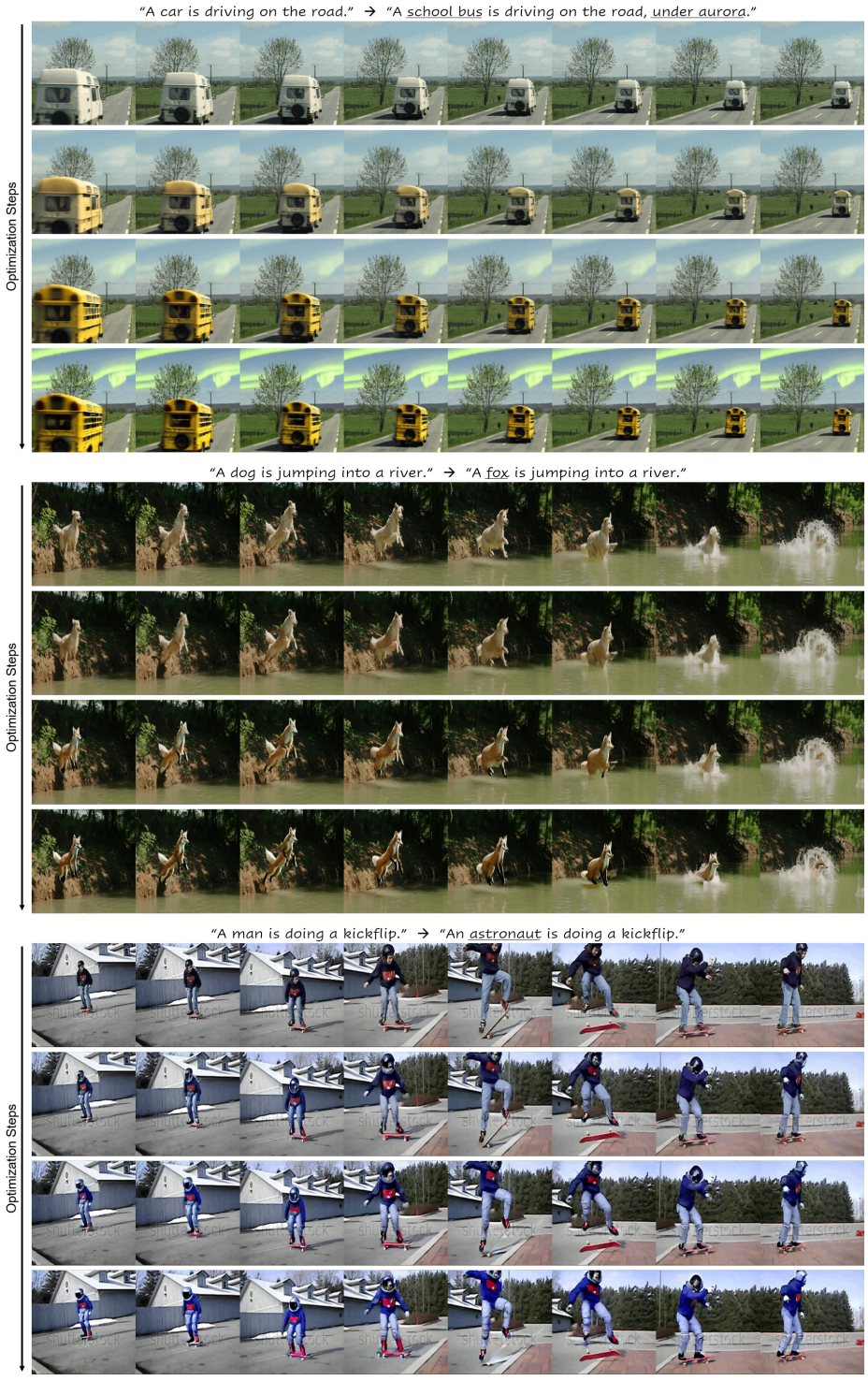

Fig. 4: Visualization of optimization progress.

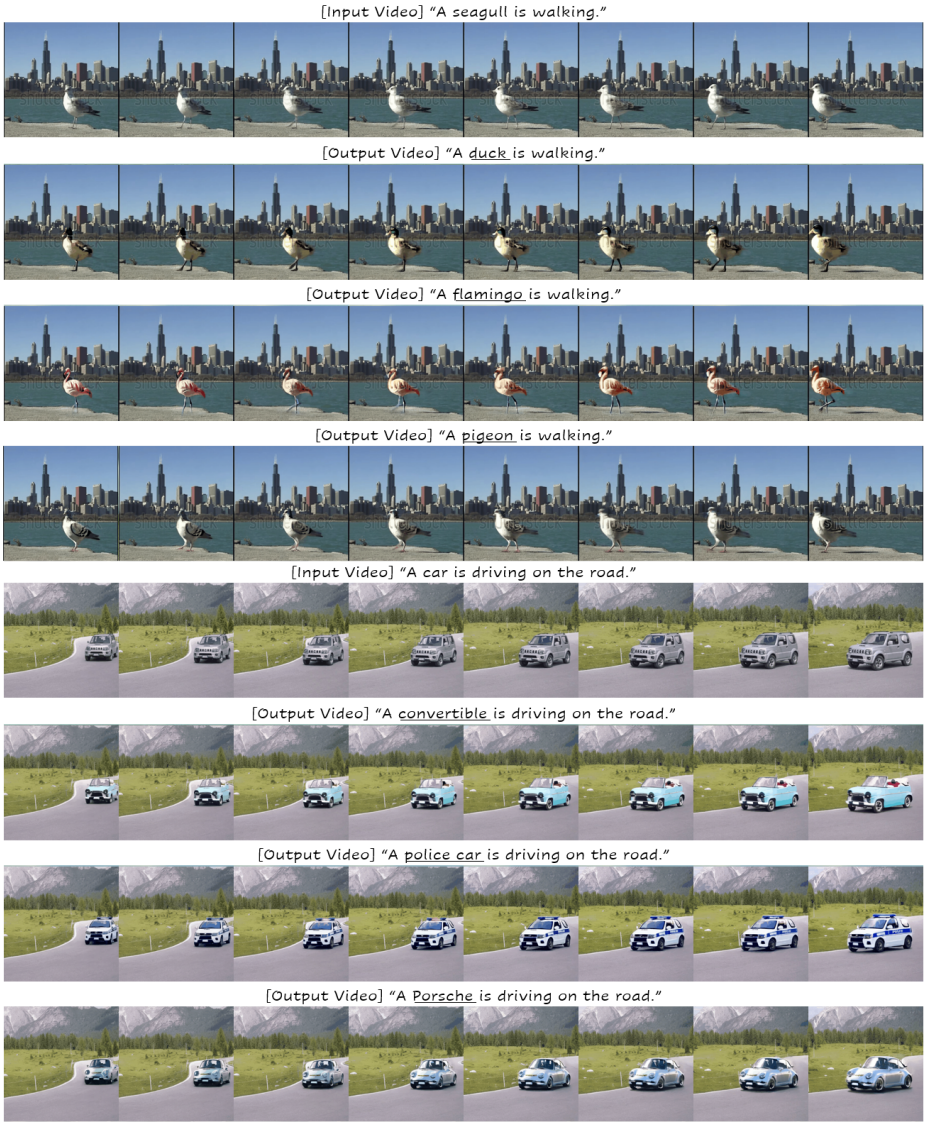

Fig. 5: Additional results of DreamMotion with Zeroscope T2V.

Fig. 6: Additional results of DreamMotion with Zeroscope T2V.

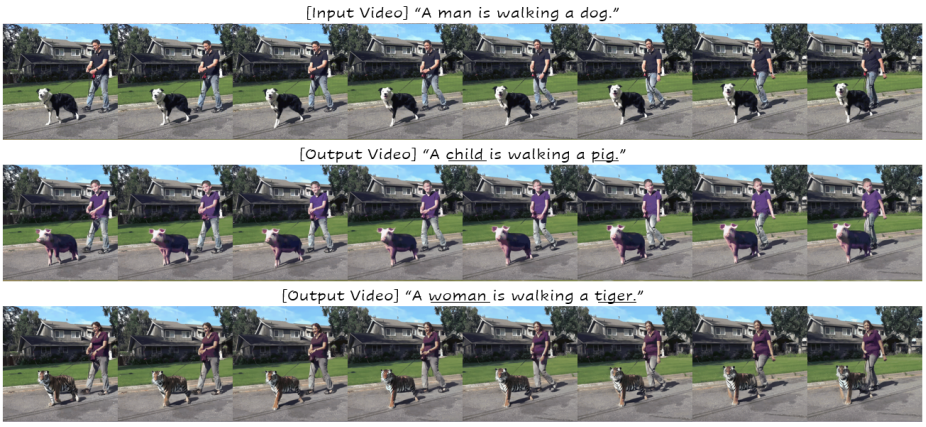

Fig. 7: Additional results of DreamMotion with Zeroscope T2V.

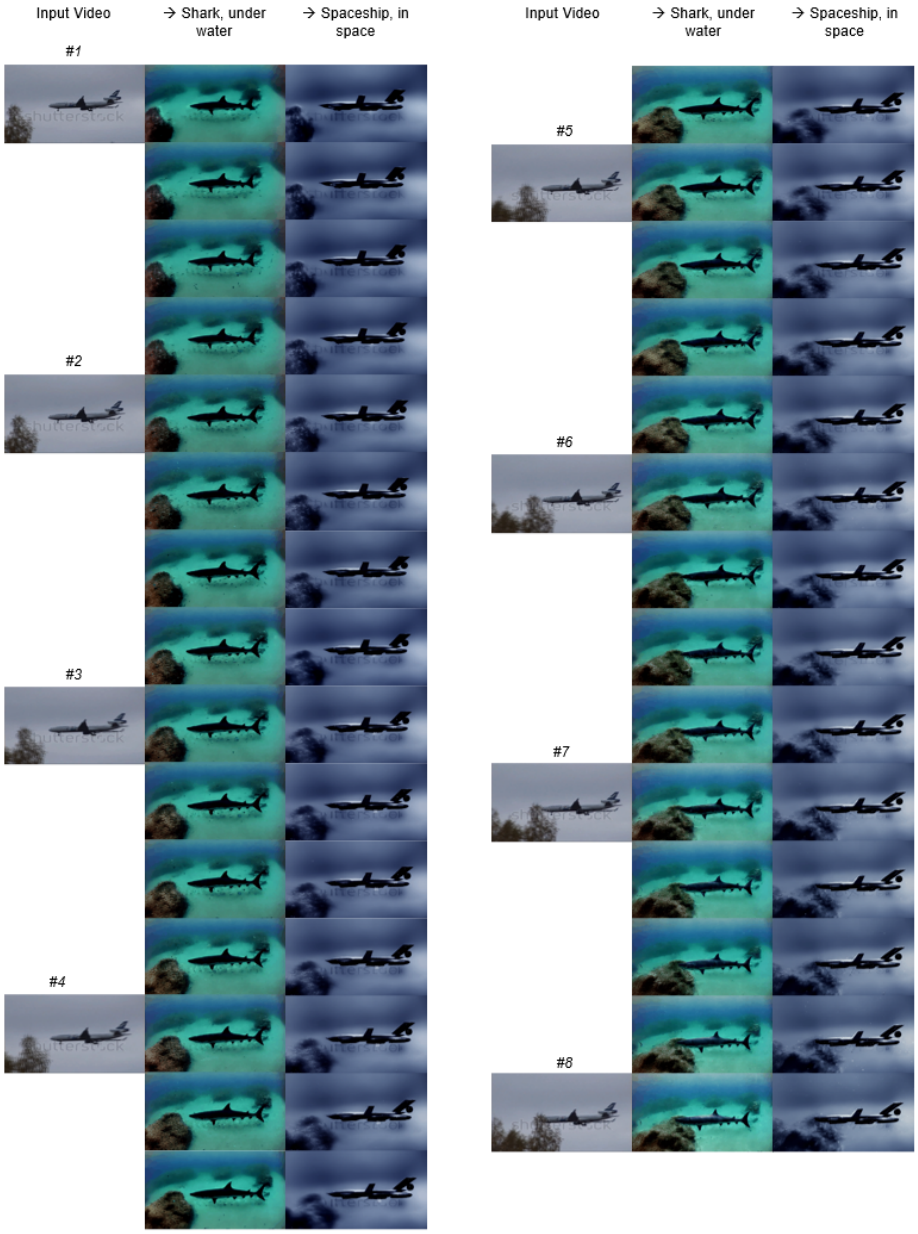

Fig. 8: Additional results of DreamMotion with Show-1 Cascade.

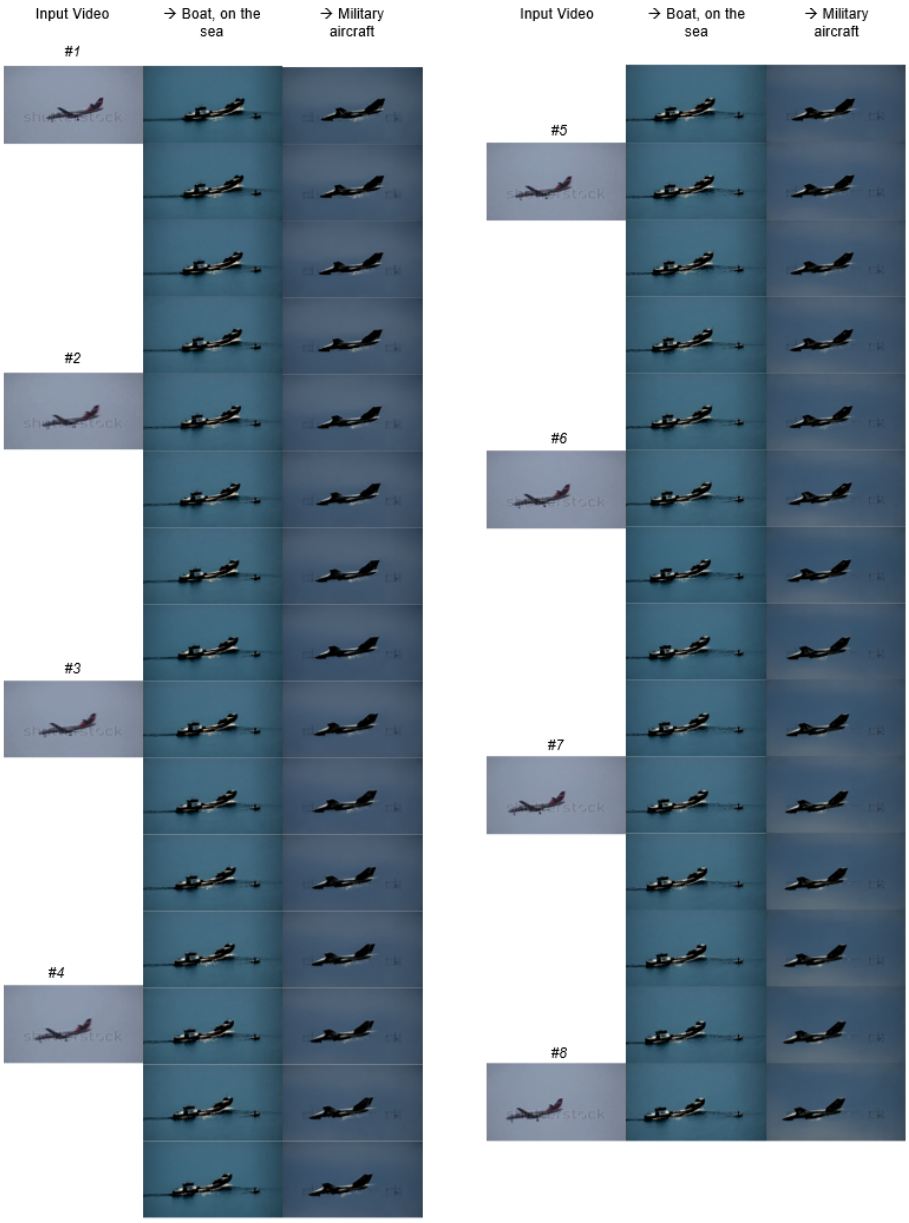

Fig. 9: Additional results of DreamMotion with Show-1 Cascade.

Fig. 10: Additional results of DreamMotion with Show-1 Cascade.

References

1. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
2. 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)
3. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
4. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
5. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
6. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
7. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023)
8. Huang, Y., Wang, J., Shi, Y., Qi, X., Zha, Z.J., Zhang, L.: Dreamtime: An improved optimization strategy for text-to-3d content creation. arXiv preprint arXiv:2306.12422 (2023)
9. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9212–9221 (2024)
10. Jeong, H., Ye, J.C.: Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In: The Twelfth International Conference on Learning Representations
11. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
12. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
13. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)
14. Materzynska, J., Sivic, J., Shechtman, E., Torralba, A., Zhang, R., Russell, B.: Customizing motion in text-to-video diffusion models. arXiv preprint arXiv:2312.04966 (2023)
15. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)

16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
17. Nam, H., Kwon, G., Park, G.Y., Ye, J.C.: Contrastive denoising score for text-guided latent diffusion image editing. arXiv preprint arXiv:2311.18608 (2023)
18. Park, G.Y., Jeong, H., Lee, S.W., Ye, J.C.: Spectral motion alignment for video motion transfer using diffusion models. arXiv preprint arXiv:2403.15249 (2024)
19. Park, J., Kwon, G., Ye, J.C.: Ed-nerf: Efficient text-guided editing of 3d scene using latent space nerf. arXiv preprint arXiv:2310.02712 (2023)
20. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
21. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)
24. Sterling, S.: Zeroscope (2023), `https://huggingface.co/cerspense/zeroscope_v2_576w`
25. Tsalicoglou, C., Manhardt, F., Tonioni, A., Niemeyer, M., Tombari, F.: Textmesh: Generation of realistic 3d meshes from text prompts. arXiv preprint arXiv:2304.12439 (2023)
26. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
27. 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)
28. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36** (2024)
29. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. arXiv preprint arXiv:2312.04433 (2023)
30. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
31. Yatim, D., Fridman, R., Tal, O.B., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. arXiv preprint arXiv:2311.17009 (2023)
32. 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. arXiv preprint arXiv:2309.15818 (2023)
33. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

34. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
35. Zhang, Y., Tang, F., Huang, N., Huang, H., Ma, C., Dong, W., Xu, C.: Motioncrafter: One-shot motion customization of diffusion models. arXiv preprint arXiv:2312.05288 (2023)
36. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: Adding conditional control for one shot text-to-video editing. arXiv preprint arXiv:2305.17098 (2023)
37. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)
38. Zhu, J., Zhuang, P.: Hifa: High-fidelity text-to-3d with advanced diffusion guidance. arXiv preprint arXiv:2305.18766 (2023)