Space-time Diffusion Features for Zero-shot Text-driven Motion Transfer Supplementary Materials

Anonymous CVPR submission

Paper ID 9898

1. Text-to-Video Model Architecture and Feature Selection

1.1. Text-to-Video Model

We use ZeroScope [1] text-to-video model, which is claimed to be fine-tuned from a Modelscope model [11] on video clips of the length of 24 frames and 576x320 resolution. Our generated results are in the same resolution with a length of 24 frames. The model was inflated from the StableDiffusion model [7] by introducing temporal layers within each building block of the UNet.

1.2. Feature Selection

The decoder of the UNet in ZeroScope comprises four blocks, each with a different resolution. We performed our analysis on coarse features, extracted from the 2nd decoder block of the UNet. We noticed that different coarse features in this block performed similarly for our task. Specifically, we tested intermediate features extracted from the spatial/temporal convolution models, output tokens from the spatial/temporal attention models, as well as features taken directly after the Upsampling block (a.k.a semantic DIFT features[9]). We empirically found that features extracted after the Upsampling block produce more visually appealing edit results.

2. Implementation Details

2.1. Feature Extraction

To obtain intermediate latents, we follow [10] and use DDIM inversion (applying DDIM sampling in reverse order) with a classifier-free guidance scale of 1 and 1000 forward steps, using a video-specific inversion prompt. We use these intermediate latents for initialization and extracting diffusion features.

2.2. Initialization and Sampling

In our experiments, we use 50 denoising steps using Restart Sampling [13] combined with DDIM sampling [8], with a classifier-free guidance scale of 10. To obtain the initial noise, we apply the downsampling/upsampling operation LF_{ξ}, described in Eq. 4 with a factor $\xi = 4$.

2.3. Optimization details

We apply the optimization described in Sec. 4.2 for the initial 20 denoising steps. In most of our experiments, we are using the Adam optimizer [5] with a learning rate of 0.01 for 30 optimization steps, but in cases where the edit required a bigger deviation from the original structure, we used a linear learning rate decay from 0.005 to 0.002 for 10 optimization steps.

2.4. Runtime

The runtime of our method mainly consists of two parts - DDIM inversion, which takes ~10 minutes, and sampling with optimization, which takes ~7 minutes for 10 optimization steps per denoising step and ~15 minutes for 30 optimization steps per denoising step, depending on the configuration.

3. Baseline Comparison Details

For comparing with Tune-A-Video [12], TokenFlow [4] and Control-A-Video [2] we used the official repositories. For visual comparison with Gen-1 [3], we used the publicly available web platform. Since this platform outputs videos of different lengths with some frames being duplicated, we excluded Gen-1 from numerical comparisons. Since SA-NLA [6] takes 10 hours to train, we compare to their provided videos and edit prompts qualitatively.

References

[1] cerspense. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 1

[2] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models, 2023. 1

[3] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1

[4] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 1

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 1

[6] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. *arXiv preprint arXiv:2301.13173*, 2023. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1

[9] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 1

[10] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to image translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 1

[11] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1

[12] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 1

[13] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi S. Jaakkola. Restart sampling for improving generative processes. *CoRR*, abs/2306.14878, 2023. 1