Supplement for "A Video is Worth 256 bases: Spatial-Temporal Expectation-Maximization Inversion for Zero-Shot Video Editing"

Maomao Li[1,2], Yu Li[2*], Tianyu Yang[2], Yunfei Liu[2], Dongxu Yue[3], Zhihui Lin[4], Dong Xu[1*]
[1]The University of Hong Kong [2]International Digital Economy Academy (IDEA)
[3]Peking University [4]Tsinghua University

https://stem-inv.github.io/page/

The supplement is organized as follows. In Sec. 1, we demonstrate that there are high correlations in video diffusion features. Then, in Sec. 2, we show more comparison between DDIM inversion and the proposed STEM inversion. Next, we display more editing results with the proposed STEM inversion in Sec. 3. Last, we show more implementation details during the editing process in Sec. 4.

1. High Correspondence in Diffusion Feature

The previous work [5] has proven that *correspondence emerges in image diffusion models (e.g., StableDiffusion) without any explicit supervision*, where diffusion features can be used to find matching pixel locations in two images by a simple nearest neighbor lookup. Fig A1 shows correspondences between video frames using diffusion features. Since there are high correlations in video diffusion features, we can use the EM algorithm to identify the low-rank representation for the entire video.

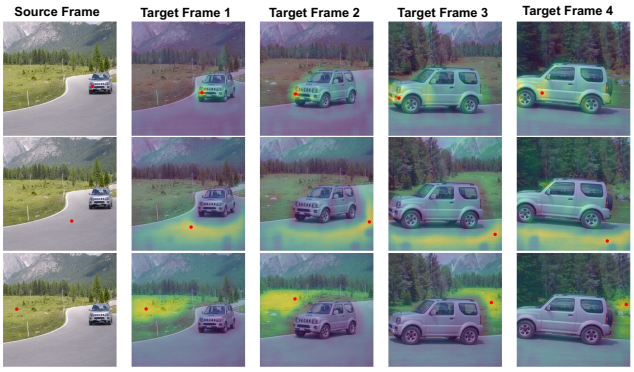

Figure A1. Given a different source pixel, the best matching pixel from the target frames can be predicted via diffusion features.

2. Inversion Comparison

2.1. DDIM Inversion Using All-frame Context

Recall that DDIM inversion in existing video editing methods [2, 4, 6] usually exploits 1-frame or 2-frame context to invert each frame. Thus, we design a more radical inflated DDIM inversion that uses all-frame context as a reference in Table 1 of our paper. Here, we use the typical DDIM reconstruction method to provide a video reconstruction comparison in Fig. A2, where both our STEM inversion and the radical inflated DDIM one can explore context from the entire video, while the resource-consuming latter yields inferior performance. The quantitative comparison in Tab. A3 also supports these findings, where we record the average PSNR and SSIM of 5 reconstruction videos. We argue that the underlying reason is those redundant or abnormal features can be effectively removed by evaluating low-rank representations.

2.2. Feature Similarity in Different Forward Steps

Recall that in Fig. 6 of our main paper, we first use optical flow to warp the former-frame features and obtain the warped feature. Then, we calculate the cosine similarity between the wrapped feature from the former frame and the current frame feature. The more similar, the reconstructed video is more coherent in time dimension.

In this supplement, we provide the mean cosine similarity across different time steps t in Fig. A3. The higher similarity indicates that our STEM inversion can achieve better temporal consistency from the perspective of optical flow.

2.3. STEM Inversion with Various Video Lengths and Video Resolutions

We provide average PSNR and SSIM between 5 reconstruction videos and ground-truth ones in Tab. A1 and Tab. A2. For Tab. A1, we sample original videos evenly whose size is 512^2, and form 8, 16, 32, 64, and 128-frame videos separately. For Tab. A2, the frame number is 24. Our

*Corresponding Author

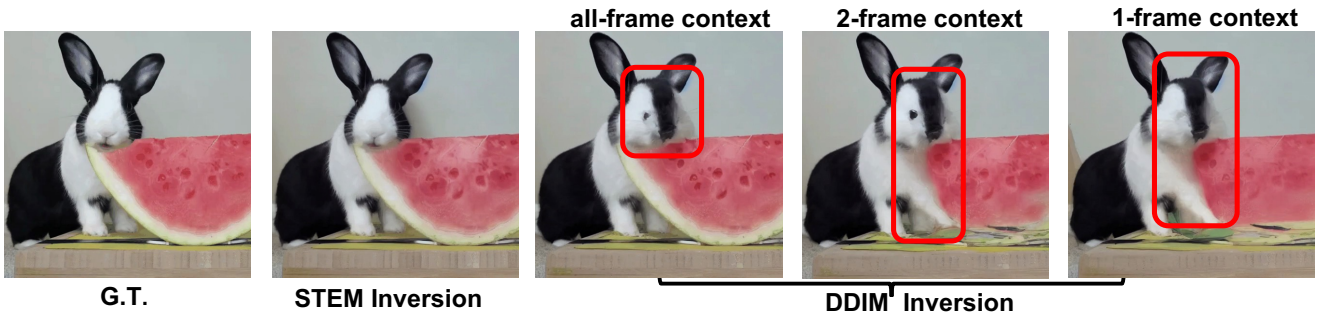

Figure A2. Reconstruction results with DDIM inversion and the proposed STEM inversion, respectively.

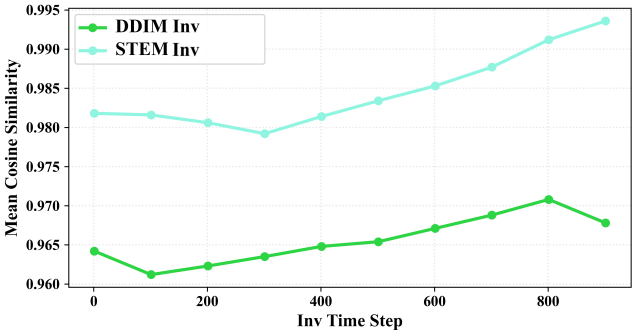

Figure A3. The mean cosine similarity between the warpped features from the former frame and the target features during various inversion steps. The more similar, the better.

Length	8	16	32	64	128
PSNR	32.702	**33.279**	33.197	33.058	32.751
SSIM	0.9836	**0.9854**	**0.9854**	0.9851	0.9843

Table A1. Results of STEM inversion with various video lengths.

Size	384^2	512^2	768^2	1024^2
PSNR	29.558	33.767	34.813	**35.668**
SSIM	0.9743	0.9847	0.9901	**0.9902**

Table A2. Results of STEM inversion with various video resolutions.

STEM inversion achieves the best reconstruction when the frame number is 16 and the resolution is 1024^2.

2.4. More Reconstruction Comparison

We provide the reconstruction comparison of DDIM inversion and STEM inversion in Fig. A4 and our project page. Since Fatezero [4] stores the intermediate self-attention maps and cross-attention maps at each timestep t, it is memory-consuming and cannot perform video editing over 20 frames on a single A100 GPU. On our project page, we sample FateZero results at a proper rate.

As seen in Fig. A4, two reconstruction fashions are applied for DDIM and STEM inversion separately: (i) the typical DDIM reconstruction (used by TokenFlow [2]), (ii) DDIM reconstruction with extra attention fusion (used by FateZero [4]). The proposed STEM inversion always delivers better reconstruction than DDIM inversion, especially under the typical DDIM reconstruction fashion. Such a benefit is derived from our STEM inversion modelling global and fixed context for each frame while DDIM inversion explores a time-varying and limited spatial-temporal context.

3. More Comparison of Various Text-driven Zero-shot Video Editing

To prove the efficiency of our STEM inversion, we compare our STEM-TokenFlow and STEM-FateZero with the current state-of-the-art video editing methods in Fig. A5 and our project page. Specifically, although T2V-Zero [3] can perceive the style and subject to be edited, it always deviates greatly from the original video and cannot maintain a satisfying temporal consistency. Besides, Tune-A-Video [6] needs to perform training on the video before editing while its performance is inferior to ours. Moreover, it is difficult for Pix2Video [1] to maintain the background. Please see the second from last row of Fig. A5.

Note that FateZero [4] struggles to conduct shape editing (see "cow" $\rightarrow$ "boar"). Our STEM inversion is able to endow shape-editing ability to FateZero, which also demonstrates the superiority of our method. Besides, by replacing DDIM inversion with the proposed STEM inversion, TokenFlow also yields more high-quality video editing results. As seen in Fig. A5, our STEM-TokenFlow has better editing fidelity when transferring the video to *Johannes Vermeer style*.

4. More Implementation Details

The attention fusion ratio in FateZero [4] is a hyperparameter controlling the editing effect. Specifically, it fuses both cross-attention and self-attention at DDIM time step $t \in [0.5T, T]$. However, we experimentally discovered a

DDIM Inv (1-frame)		DDIM Inv (2-frame)		DDIM Inv (all-frame)		STEM Inv	
PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
24.122	0.8137	25.967	0.8595	26.464	0.8700	**31.572**	**0.9606**

Table A3. Qualitative comparison between different inversion. Here, "1-frame", "2-frame", and "all-frame" refer to the context frames considered during single-frame inversion calculation for DDIM inversion.

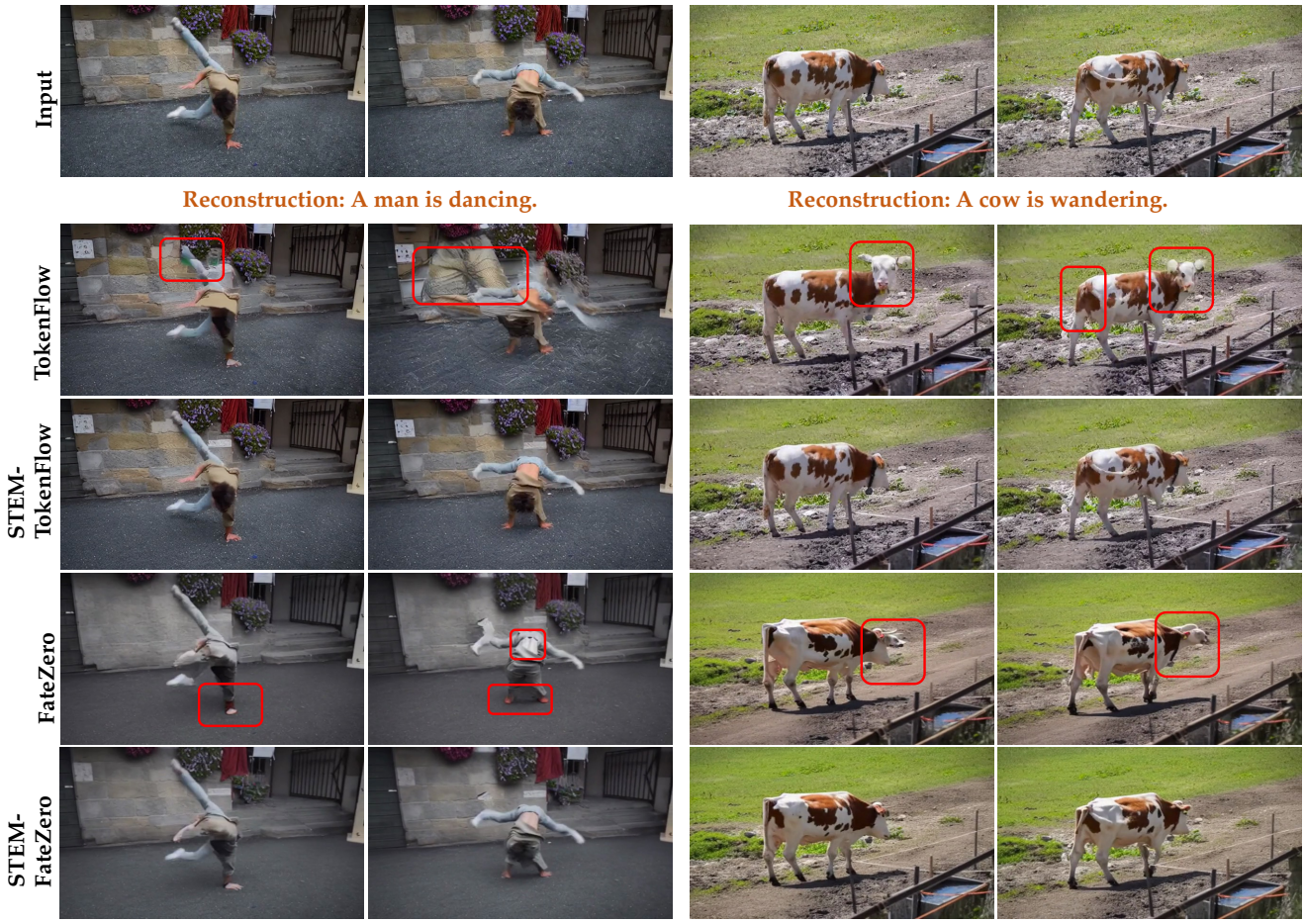

Figure A4. Qualitative comparison of the reconstruction with DDIM and STEM inversion, where two reconstruction fashions are applied: (i) DDIM reconstruction (i.e., TokenFlow [2] reconstruction)), (ii) DDIM reconstruction with additional attention fusion (i.e., Fatezero [4] reconstruction).

small fusion ratio for cross attention time step is better when using our STEM inversion for Fatezero editing. Concretely, we adopt the cross attention time step ratio as $t \in [0.2T, T]$, while the same ratio $t \in [0.5T, T]$ for self-attention. The possible underlying reason is that our STEM inversion is more sensitive to capturing the semantics from the target prompt than the DDIM one. Thus, a smaller cross-attention fusion ratio is sufficient under the FateZero editing scenario. Besides, in terms of TokenFlow editing, we use identical hyper-parameters when replacing DDIM inversion with our STEM inversion.

References

[1] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 2

[2] 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, 2, 3, 4

[3] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-

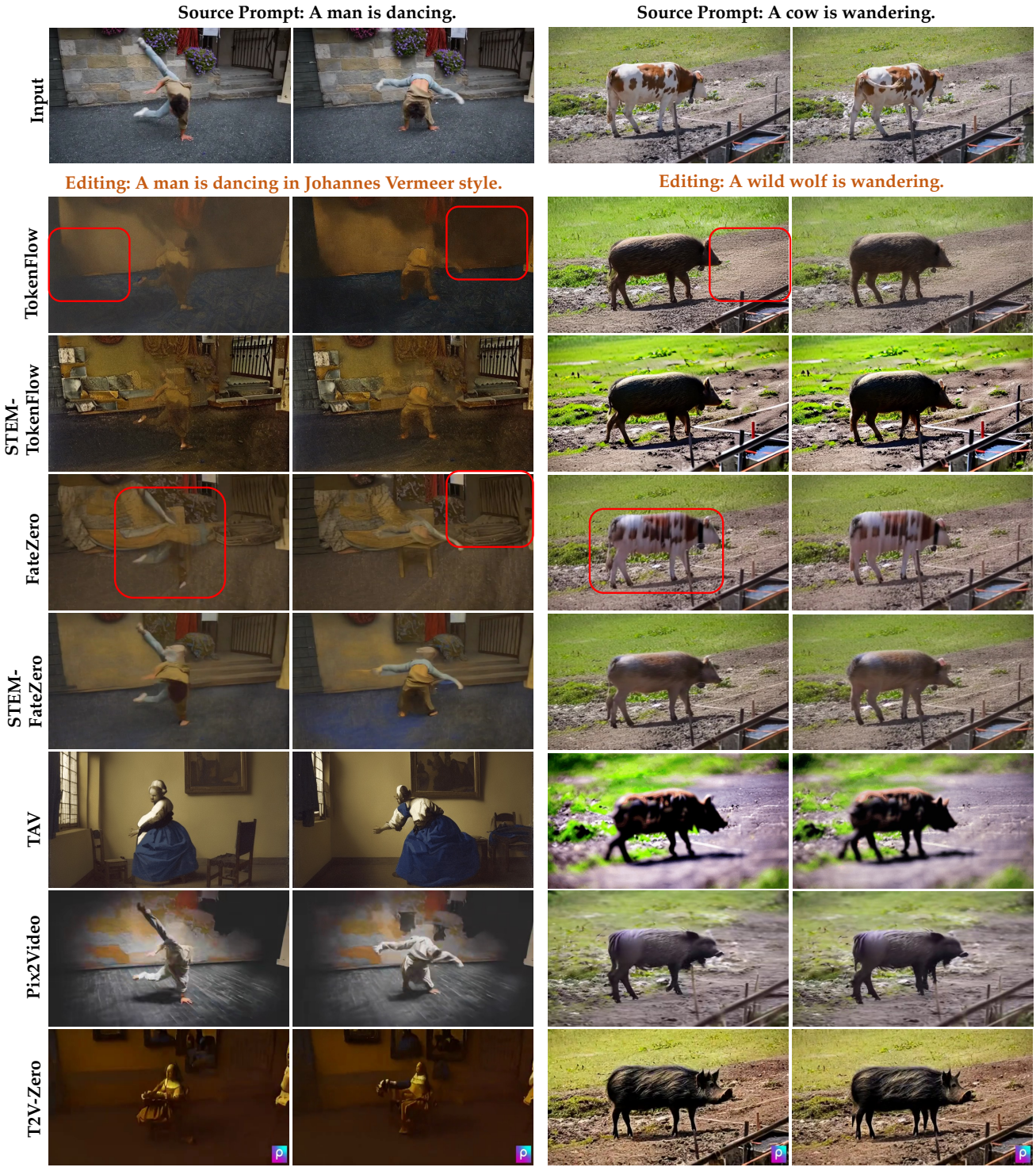

Figure A5. Qualitative comparison between different text-driven video editing methods. Our STEM-inversion can consistently improve the editing performance of TokenFlow [2] and FateZero [4]. Best viewed with zoom-in.

to-image diffusion models are zero-shot video generators. *ICCV*, 2023. 2

[4] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing

attentions for zero-shot text-based video editing. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3, 4

[5] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1

[6] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 1, 2