DIVE: Taming DINO for Subject-Driven Video Editing

Supplementary Material

1. Additional Experiments

1.1. Comparison with Other Feature Extractors

To assess the effectiveness of DINO in DIVE, we replace it with two alternative feature extractors, CLIP Image Encoder [9] and Google ViT [3], and evaluate their impact across all three stages of DIVE. As shown in Table 1, DINO consistently outperforms both alternatives in all key metrics, particularly in image alignment, temporal consistency and overall video quality, demonstrating superior identity preservation and motion coherence. The primary reason for CLIP and ViT's inferior performance lies in their weaker ability to capture fine-grained semantic correspondences. CLIP is optimized for image-text alignment and lacks strong spatial discrimination, making it less effective at distinguishing subject details across frames. Google ViT, although designed for general vision tasks, does not explicitly focus on semantic consistency, leading to suboptimal identity preservation. In contrast, DINO features exhibit strong semantic consistency across frames while maintaining precise part-level discrimination, ensuring accurate motion alignment and subject identity retention. This highlights DINO's advantage in subject-driven video editing.

Feature Extractor	Text Alignment [11] ↑	Image Alignment [7] ↑	Temporal Consistency [11] ↑	Overall Video Quality [6] ↑
	Reference Image Guided Subject Editing			
CLIP Image Encoder	29.10	81.34	89.20	0.591
Google ViT	28.44	77.25	89.41	0.437
DINO	**29.43**	**84.27**	**92.33**	**0.775**
	Text Guided Subject Editinge			
CLIP Image Encoder	31.26	\	91.17	0.539
Google ViT	30.76	\	90.31	0.502
DINO	**32.29**	\	**95.89**	**0.614**

Table 1. Quantitative ablations of feature extractors.

1.2. Compare to Object Tracking Methods

We compared our DINO-based motion guidance with three alternatives used in object tracking: optical flow [2], segmentation masks [1], and Stable Diffusion features [12]. DINO features were replaced only in the first stage while other stages remained unchanged. As shown in the following table, our method achieves higher temporal consistency and alignment, demonstrating the effectiveness of DINO as semantically robust motion guidance.

Method	Temporal Consistency ↑	Text Alignment↑	Image Alignment ↑
optical flow [2]	88.24	28.49	69.58
masks [1]	83.54	28.35	63.29
Stable Diffusion features [12]	90.35	28.67	81.22
DINOv2 features (Ours)	**92.33**	**29.43**	**84.27**

Table 2. Quantitative ablations of object tracking methods.

1.3. Generalization to Stronger Backbone

DIVE is model-agnostic and can be applied with various diffusion backbones. Applied to SDXL, it shows consistent improvements over SD 1.5 (see below).

Backbone	Text Alignment ↑	Image Alignment ↑	Temporal Consistency ↑
SDXL	+9.84%	+14.89%	+20.03%

Table 3. Quantitative ablations of different backbones.

1.4. Comparison with Other Methods

In addition to the four state-of-the-art video editing methods compared in the main paper, we further evaluate DIVE against two recent approaches: TokenFlow [5] and I2VEdit [8]. The qualitative comparison are presented in the project page[1], which provides a clearer visualization. Compared to these methods, DIVE achieves more precise subject identity preservation and better motion alignment, benefiting from its dedicated identity registration and motion modeling stages. This highlights the trade-off between zero-shot efficiency and fine-tuned accuracy in subject-driven video editing.

2. Implementation of Learnable MLPs

In the first and second stages of DIVE, we incorporate the extracted DINO features into the diffusion space using four learnable MLPs, $\psi = \{\psi_l | l \in \{1, 2, 3, 4\}\}$ and $\phi = \{\phi_l | l \in \{1, 2, 3, 4\}\}$, respectively. This projection process is illustrated in Figure 1. Here, H and W denote the height and width of each latent variable transformed

[1] https://dino-video-editing.github.io/

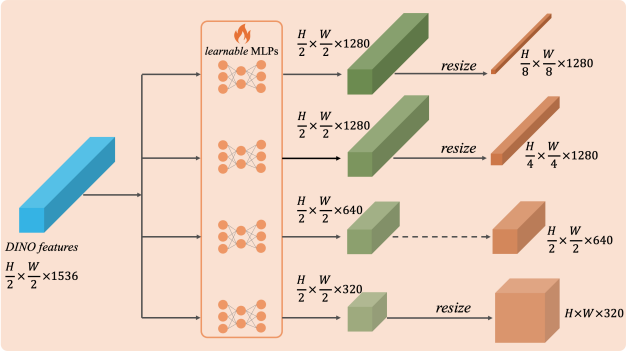

Figure 1. Visualization of the projection process via learnable MLPs in the first and second stages. Note that in implementation, the tensor shape is $B \times N \times H \times W \times C$ with B and N being the batch size and the number of frames; for simplicity, B and N are omitted there.

from a single frame or image. We use the ViT-g/14 variant without registers[2] as the DINOv2 backbone, with the channel dimension of 1536. Correspondingly, the channel dimensions of the intermediate features after each downsample block in the Stable Diffusion U-Net encoder are $\{320, 640, 1280, 1280\}$.

3. Limitations

3.1. Identity Inconsistency

Similar to previous image personalization methods like DreamBooth [10] and Textual Inversion [4], DIVE may also occasionally produce artifacts in preserving the target subject's identity, as illustrated in the first row of Figure 2. Additionally, the identity across the edited video frames may be inconsistent in complex scenes involving intricate interactions, occlusions, or significant view changes, as shown in the second row of Figure 2. These challenges primarily result from the limited generative capacity of the pretrained base model. Future work could explore integrating DIVE with more robust text-to-video models to improve generalization and identity consistency.

3.2. Time Cost

Another limitation of DIVE is the computational overhead due to its reliance on testing-time fine-tuning. Specifically, the total time required for editing a video with specific reference images is approximately 10 minutes, broken down as follows: 2 minutes for stage 1, 8 minutes for stage 2, and 30 seconds for stage 3. Future work could focus on developing zero-shot editing capabilities with DINO to improve efficiency.

[2] https://github.com/facebookresearch/dinov2

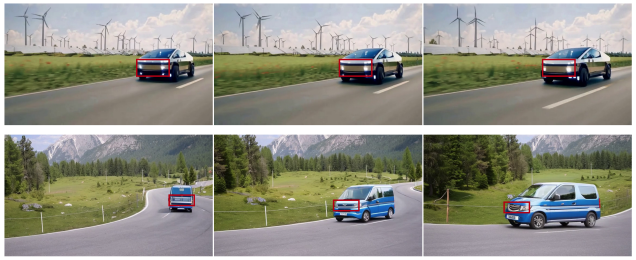

Figure 2. Examples of DIVE's limitations in identity preservation and consistency.

References

[1] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 1

[2] Seokju Cho, Jiahui Huang, Seungryong Kim, and Joon-Young Lee. Flowtrack: Revisiting optical flow for long-range dense tracking. In *CVPR*, pages 19268–19277, 2024. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[5] 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

[6] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024. 1

[7] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 1

[8] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[10] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2

[11] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei

Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023. 1

[12] Zhengbo Zhang, Li Xu, Duo Peng, Hossein Rahmani, and Jun Liu. Diff-tracker: text-to-image diffusion models are unsupervised trackers. In *ECCV*, pages 319–337. Springer, 2024. 1