
DFVEdit: Conditional Delta Flow Vector for Zero-shot Video Editing

Anonymous Author(s)

Affiliation

Address

email

1 A Additional theoretical details

2 A.1 Revisiting video editing from sampling perspective

3 Let $\{X_t^{\text{edit}}\}_{t=0}^T$ define the state trajectory of the edited video in the sampling process. We formalize
4 video editing as a *controlled Markov chain* with the following recursive relation:

$$X_{t-1}^{\text{edit}} = g_{\theta_2,t} \left(X_t^{\text{edit}}, \underbrace{\epsilon_{\theta_1}(X_t^{\text{edit}}, t)}_{\text{Canonical Denoiser}} + \lambda \underbrace{C(X_t^{\text{edit}}, t, *)}_{\text{Control Term}} \right) \quad (1)$$

5 where *State Transition* $g_{\theta_2,t}$ is the differentiable transition function parameterized by learnable θ_2 ,
6 ϵ_{θ_1} is the pretrained diffusion model with frozen θ_1 , *Control Term* C is the editing condition injector
7 with intensity $\lambda \geq 0$.

8 The formulation maintains *consistency with standard diffusion sampling process* when $\lambda = 0$ and
9 $g_{\theta_2} = \mathcal{I}$, where $\mathcal{I} : \mathcal{X} \rightarrow \mathcal{X}$ denotes the identity operator satisfying $\mathcal{I}(x) = x, \forall x \in \mathcal{X}$:

$$X_{t-1}^{\text{edit}} \Big|_{\substack{\lambda=0 \\ g_{\theta_2}=\mathcal{I}}} \equiv X_{t-1}^{\text{orig}} \quad (2)$$

10 A.1.1 Unification with various editing methods

11 Existing popular editing paradigms emerge as special cases of our control framework:

12 1. **Inversion-based editing** (like Fatezero [1]):

$$g_{\theta_2,t}(a, b) = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} (a + \Delta\beta_t b) \quad (3)$$

$$C(X_t^{\text{edit}}, t, *) = \epsilon_{\theta_1}^{\text{edit}}(X_t^{\text{edit}}, t) - \epsilon_{\theta_1}(X_t^{\text{edit}}, t) \quad (4)$$

$$\Delta\beta_t = \sqrt{\frac{1 - \alpha_{t-1}}{\alpha_{t-1}}} - \sqrt{\frac{1 - \alpha_t}{\alpha_t}} \quad (5)$$

13 2. **Latent-approximation-based editing** (like DDS [2]):

$$g_{\theta_2,t}(a, b) = \text{Proj}_{\theta_2,t}(a + \eta b) \quad (6)$$

$$C(x_t, t, *) = \epsilon_{\theta_1}(x_t, t) - \epsilon_{\theta_1}(x_t, t) - \epsilon \quad (7)$$

$$\epsilon \sim \mathcal{N}(0, \sigma_t^2 I) \quad (8)$$

where α_t is the DDPM noise schedule coefficient at step t , $\Delta\beta_t$ is the noise scale difference term maintaining consistency in the reverse process, $\epsilon_{\theta_1}^{\text{edit}}$ is the edited noise prediction conditioned on the target prompt, X_t^{edit} is the latent representation during the editing process. And $Proj_{\theta_2, t}$ is a shallow approximation network with learnable parameter θ_2 that directly refines the latent to the target latent, η is the step size controlling parameter update strength, σ_t is the time-dependent noise scale for stochastic refinement, and ϵ is the Gaussian noise enabling exploration in the latent space. These formulations show how various editing methods are implicitly isomorphic with the sampling process.

A.1.2 Revisiting video editing from the continuous flow transformation perspective

For the sampling process of SDE:

$$dX = \underbrace{-\frac{1}{2}\beta(t)X}_{f(X,t)} dt + \underbrace{\sqrt{\beta(t)} dW}_{g(t)} \quad (9)$$

For the inverse process of SDE:

$$dX = \left[-\frac{1}{2}\beta(t)X - \beta(t)\nabla_X \log p_t(X) \right] dt + \sqrt{\beta(t)}dW \quad (10)$$

when changing the discrete update formulation into continuous $\Delta t \rightarrow 0$, we define:

$$\alpha(t) = e^{-\int_0^t \beta(s)ds}, \quad \sigma(t) = \sqrt{\frac{1 - \alpha(t)}{\alpha(t)}} \quad (11)$$

Using Taylor's expansion, we have:

$$\alpha_{t-1} \approx \alpha(t) - \dot{\alpha}(t)\Delta t \quad (12)$$

$$\frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} \approx 1 - \frac{1}{2}\beta(t)\Delta t \quad (13)$$

$$X_{t-1}^{\text{edit}} \approx X_t^{\text{edit}} - \frac{\beta(t)}{2} (X_t^{\text{edit}} + \sigma(t)(\epsilon_{\theta_3} + \lambda C)) \Delta t \quad (14)$$

Under the Euler discretization scheme with step size $\Delta t \rightarrow 0$ and $g_{\theta_2} = \mathcal{I}$, the discrete process (1) converges to the controlled SDE:

$$dX_t^{\text{edit}} = \underbrace{\left[-\frac{\beta(t)}{2}X_t^{\text{edit}} + \frac{\beta(t)}{2}\nabla \log p_t(X_t^{\text{edit}}) + \lambda \frac{\beta(t)}{2}\sigma(t)C(X_t^{\text{edit}}, t) \right]}_{f_{\theta_1}(X_t^{\text{edit}}, t)} dt + \underbrace{\sqrt{\beta(t)} dW}_{g(t)} \quad (15)$$

And our derived CDFV adheres to the *minimum intervention principle* from optimal control theory, which theoretically guarantees computational efficiency:

$$\min_{\lambda, C} \mathbb{E} \left[\int_0^T \|C(X_t, t)\|^2 dt \right] \quad \text{s.t.} \quad dX = [f_{\theta_1}(X, t) + \lambda C(X, t)] dt + g(t)dW \quad (16)$$

$$C^*(X, t) = \frac{\nabla_X \log p_t^{\text{edit}}(X) - \nabla_X \log p_t(X)}{\sigma(t)} \quad (17)$$

In addition, we provide the simplified algorithm of DFVEdit as below:

Algorithm 1 Simplified algorithm for DFVEdit

Require: source video \mathbf{X}_0 , target and source prompt embeddings $[C_1, C_0]$, Video DiT ϵ_{θ_1} , encoder $\mathcal{E}(\cdot)$, decoder $\mathcal{D}(\cdot)$, sampling timesteps T , ER scale $\gamma^{(k)}$

Ensure: Edited video \mathbf{X}_1

```
1:  $\mathbf{Z}_0 \leftarrow \mathcal{E}(\mathbf{X}_0)$  ▷ Latent encoding
2:  $\hat{\mathbf{Z}}_T \leftarrow \mathbf{Z}_0$  ▷ Initialize target latent
3:  $\tilde{C}_1 \leftarrow C_1 + \gamma^{(k)} \odot C_1$  ▷ Embedding Reinforcement
4: for  $t \leftarrow T$  down to 1 do
5:    $\mathbf{Z}_{\text{trans}} \leftarrow \Phi_t([\hat{\mathbf{Z}}_t; \mathbf{Z}_0])$  ▷ One-step forward process:  $q(\mathbf{z}_t|\mathbf{z}_0)$ 
6:    $\Delta v_t \leftarrow \frac{v_{(t, c_1)}, v_{(t, c_0)}}{\Delta} \epsilon_{\theta_1}(\mathbf{Z}_{\text{trans}}, [\tilde{C}_1, C_0])$  ▷ Raw CDFV prediction
7:    $\Delta v_{(t, M_t)} \leftarrow M_t \odot [\Delta v_t]$  ▷ Implicit Cross-Attention Guidance
8:    $\hat{\mathbf{Z}}_{t-1} \leftarrow \hat{\mathbf{Z}}_t - \Delta v_{(t, M_t)}$  ▷ Latent update
9: end for
10:  $\mathbf{X}_1 \leftarrow \mathcal{D}(\hat{\mathbf{Z}}_0)$  ▷ Video synthesis
```

Note:

- ▷ Flow map Φ_t implements the one-step forward process with standard method-specific coefficients (DDPM/DDIM/Flow Matching).
 - ▷ ICA mask M_t is computed from the specific layer of the Full Attention map (Section 3.3).
-

31 B Additional experimental details

32 B.1 Experimental settings

33 We do quantitative and human evaluations with 8 quantitative metrics, including Temporal Consistency
34 ('CLIP-F'), Warping Error (' E_{warp} ') [3], Prompt Alignment ('CLIP-T'), Masked PSNR ('M.PSNR'),
35 Perceptual Similarity ('LPIPS'), Relative GPU Memory Consumption ('VRAM'), Relative CPU
36 Memory Consumption ('RAM'), Relative Inference Latency ('Latency') and 3 metrics for user study,
37 including Text Alignment ('Edit'), Overall Frame Quality ('Quality') and Temporal Consistency and
38 Realism ('Consistency').

39 **CLIP metrics.** We employ the output logits of the official ViT-L-14 CLIP model to compute two
40 metrics: (1) the mean cosine similarity between all frame embeddings and the target text prompt
41 (CLIP-T), and (2) the average cosine similarity of consecutive frame embeddings of edited videos
42 (CLIP-F).

43 **Masked PSNR.** We quantify structural preservation by computing Masked PSNR on unedited regions,
44 following [4]. Using 10 DAVIS [5] videos, 30 diverse prompts and corresponding segmentation
45 annotations M provided by DAVIS, we calculate pixel-level differences between source (\mathbf{X}_0) and
46 edited (\mathbf{X}_1) videos within regions identified by inverted segmentation masks $M^* = \neg M$.

$$M.PSNR(\mathbf{X}_1, \mathbf{X}_0) = PSNR(B(\mathbf{X}_1, M^*), B(\mathbf{X}_0, M^*)) \quad (18)$$

47 where $B(\dots)$ is the binary operation with a threshold of 0.3.

48 **User study.** We conducted a pairwise comparison study with 20 participants evaluating 80 video-
49 prompt pairs (30 from DAVIS, 50 from the website Pexels [6]). Participants rated three aspects:
50 (1) *Text Alignment* (prompt-video correspondence), (2) *Frame Quality* (visual artifacts), and (3)
51 *Consistency* (temporal coherence and motion preservation). Scores (0-100 scale) were aggregated by
52 trimming extremes and averaging remaining responses, yielding 1600 total ratings.

53 **Relative Efficiency Metrics.** We evaluate computational efficiency through three normalized metrics:
54 Relative CPU Memory Consumption means the ratio of average CPU Memory allocated during
55 editing to that average CPU allocated to original inference (only generation) with the same base
56 model. Relative Inference Latency means the ratio of the latency with editing to that of the original
57 inference latency with the same model. These relative metrics enable fair comparison across varying
58 base model requirements. Tab. T1 reports absolute values and experimental configurations. Reported
59 values include absolute peak allocated GPU/CPU memory consumption (MB), processing latency,
60 frame count (F), and corresponding base models. The groups 'Stable Diffusion', 'Zeroscope', and

61 'CogvideoX' represent the original generation results with base models, while other groups are the
62 editing results.

Table T1: **Absolute empirical computational efficiency results.**

Method	GPU Memory (MB)	CPU Memory (MB)	Latency (s)	F	Base Model
Stable Diffusion [7]	4134.31	2865.00	15.47	8	Stable Diffusion 1.5
Zeroscope [8]	4551.32	3833.86	13.14	8	Zeroscope
CogVideoX [9]	33110.36	10522.26	100.80	41	CogVideoX-5B
SDEdit [10]	33441.46	11890.15	87.69	41	CogVideoX-5B
FateZero [1]	9576.13	61416.08	52.58	8	Stable Diffusion 1.5
FreeMask [11]	7464.16	98059.40	74.29	8	Zeroscope
TokenFlow [3]	17040.29	159475.48	126.87	8	Stable Diffusion 1.5
VideoDirector [12]	24789.38	6475.66	432.64	8	Stable Diffusion 1.5
FLATTEN [13]	6385.01	20936.98	71.35	8	Stable Diffusion 1.5
ControlVideo [14]	36115.23	4635.36	146.19	8	Stable Diffusion 1.5
DMT [15]	42500.24	25572.34	217.54	8	Stable Diffusion 1.5
Ours	31454.84	9049.14	120.96	41	CogVideoX-5B

63 B.2 Experimental details on insight

64 In Fig. 1 of the main text, we present some visualizations of our insights: (a) presents the theoretical
65 attention memory consumption of different base models; (b) presents the theoretical inference latency
66 of FateZero [1] and KVEdit [16] when directly applying them to CogVideoX-5B, as well as the
67 practical inference latency of DFVEdit on CogVideoX-5B. Here, we provide more details on these
68 insights.

69 **Attention memory explosion in DiT models.** We analyze the memory consumption of attention
70 mechanisms in the Unet module of diffusion models versus the Transformer module of DiT models,
71 focusing on estimating the memory (GB) needed for storing attention score maps in float32 format
72 per timestep. Although attention score maps are rarely computed explicitly in base models due to
73 efficiency concerns, traditional editing methods often require their direct manipulation for attention
74 engineering. Therefore, explicit examination of these maps helps in identifying challenges in
75 adapting traditional editing techniques to modern DiT architectures. While our analysis centers on
76 the memory footprint of attention score maps within a single timestep, editing methods based on
77 attention engineering may involve caching or modifying attention maps across multiple timesteps,
78 and we highlight the significant computation overhead when applying attention-engineering-based
79 video editing methods to Video DiTs. As shown in Table T2, traditional diffusion models (e.g.,
80 Stable Diffusion and Zeroscope) exhibit multi-scale attention mechanisms with shapes varying
81 by layer. In contrast, modern Video DiTs like CogVideoX-5B employ fixed large-scale attention
82 ([2, 48, 11490, 11490]), resulting in $283\times$ higher memory than SD's maximum (7 GB vs. 1871 GB).
83 This fundamental architectural shift explains the inefficiency of attention-based editing methods when
applied to Video DiTs.

Table T2: **Peak attention memory consumption (GB) for full score maps (float32) per timestep.**
Values with \sim approximate ground truth with $\pm 5\%$ variance. F : processed frames. Asterisk (*)
indicates dynamic attention shapes.

Model	Attention Shape	Block Number	Attention Memory (GB)	Dtype	F
Stable Diffusion [7]*	Multi-scale	32	~ 7	float32	1
HunyuanDiT [17]	[2,4096,4096]	80	~ 10	float32	1
Zeroscope [8]*	Multi-scale	64	~ 25	float32	8
HunyuanVideo [18]	[1,24,11520,11520]	48	~ 612	float32	41
Wanx2.1-14B [19]	[1,40,11264,11264]	40	~ 794	float32	41
CogVideoX-5B [9]	[2,48,11490,11490]	40	~ 1871	float32	41

84

85 **Inference Latency Comparison.** We adopt a theoretical estimation approach to evaluate the inference
86 latency for attention-engineering-based editing methods (FateZero [1] and KVEdit [16]) and measure
87 the practical inference latency of DFVEdit, since direct empirical testing of FateZero and KVEdit
88 with Video DiT is infeasible due to GPU memory and CPU RAM constraints. First, we conduct
89 performance analysis assuming unlimited CPU memory. Specifically, we measure execution times

per timestep and extrapolate to the total timesteps required for editing. Given that caching attention maps, keys, and queries exceeds GPU capacity, all caching operations utilize CPU memory. This methodology provides theoretical latency estimates for these methods in Video DiT contexts. The selected approaches demonstrate that both traditional diffusion-based methods and image DiT-based methods face significant resource overheads when directly applied to Video DiTs.

B.3 More experiment results

Embedding Reinforcement. Due to space limitation, we only visualize the embedding reinforcement ablation results on stylization in the main text, here we additionally visualize the results on shape editing. As shown in Fig. F1, at $\gamma = 0$, the distinctive traits of polar bears compared to brown bears, such as their white fur and rounded ears, are effectively captured with high background fidelity. Increasing γ to 1 enhances editing quality, more accurately portraying the polar bear’s elongated neck and smaller head-to-body ratio. However, at $\gamma = 5$, there is a notable decline in video synthesis quality, characterized by visible flickering and noise, alongside reduced background preservation. Our experiments demonstrate that for optimal editing outcomes in shape modification, the ER method’s hyperparameter γ should be set within the range of 0 to 1. For simplicity and efficiency, we typically set γ to 0.3 in our studies, although values within this range generally yield satisfactory results.

On multi-objects editing. Our method, though not specifically designed for multi-object editing, inherently adapts to such tasks through the editing region localization capability of CDFV. We enhance the editing precision by combining: (1) Implicit Cross-Attention (ICA) derived from Layer 16 transformer blocks (based on the observation that cross-attention masks exhibit a coarse-to-fine change across denoising timesteps as found in FreeMask [11], with the layer index selection method also following FreeMask), and (2) SAM masks with edge padding. Fig. F2 shows an example of ICA extraction and visualization. This strategy operates in two phases: ICA guidance during early denoising ($t = T \rightarrow 0.4T$) preserves shape flexibility while reducing background leakage, followed by SAM-based mask guidance ($t = 0.3T \rightarrow 0$).

As shown in Fig. F3, our method demonstrates robust multi-object editing capabilities across diverse scenarios, achieving target object accuracy while maintaining non-edited region fidelity. The framework handles both complex dynamic interactions with multiple objects in cluttered environments, and fine-grained editing requiring precise motion retention. These findings not only underscore the robustness and effectiveness of our proposed method but also lay a solid foundation for future advancements in multi-object editing.

More results on attribute editing. Due to space limitations, we include the visualization results of attribute editing in the Appendix. As shown in Fig. F4, our method demonstrates satisfactory performance on attribute editing, enabling the natural and seamless integration of both added and removed small objects within existing scenes.

More results on extension experiments. We have demonstrated additional results of applying DFVEdit to the Wan2.1 base model in the main text. To further objectively evaluate the generality and robustness of DFVEdit on Video DiTs, Figure F5 compares the performance of the same video editing tasks across different base models using our method. This comparison aims to reveal the variations in outcomes due to differences in models and to verify the robustness and generality of our approach. The experimental results indicate that although different base models may lead to slight variations in the editing outcomes, overall, all edited videos align well with the target prompt, and the editing quality meets the expected standards. These findings reflect the high robustness and generalization ability of our proposed method. *Refer to the 'DFVEdit.mp4' in the supplementary material for the dynamic video display. The code will be public upon publication of this work.*

C Limitations

Fig. F6 indicates limitations of our method. In zero-shot video editing, it is challenging to maintain full detail fidelity in non-edited regions. Although our method outperforms others in improving fidelity, achieving perfect detail fidelity remains difficult. Even with ICA guidance, some detail loss occurs. Additionally, for editing traits with large variations, such as transforming a bicycle into a car, our method cannot support significant shape variations. It is suitable only for shape editing tasks with small layout variations.

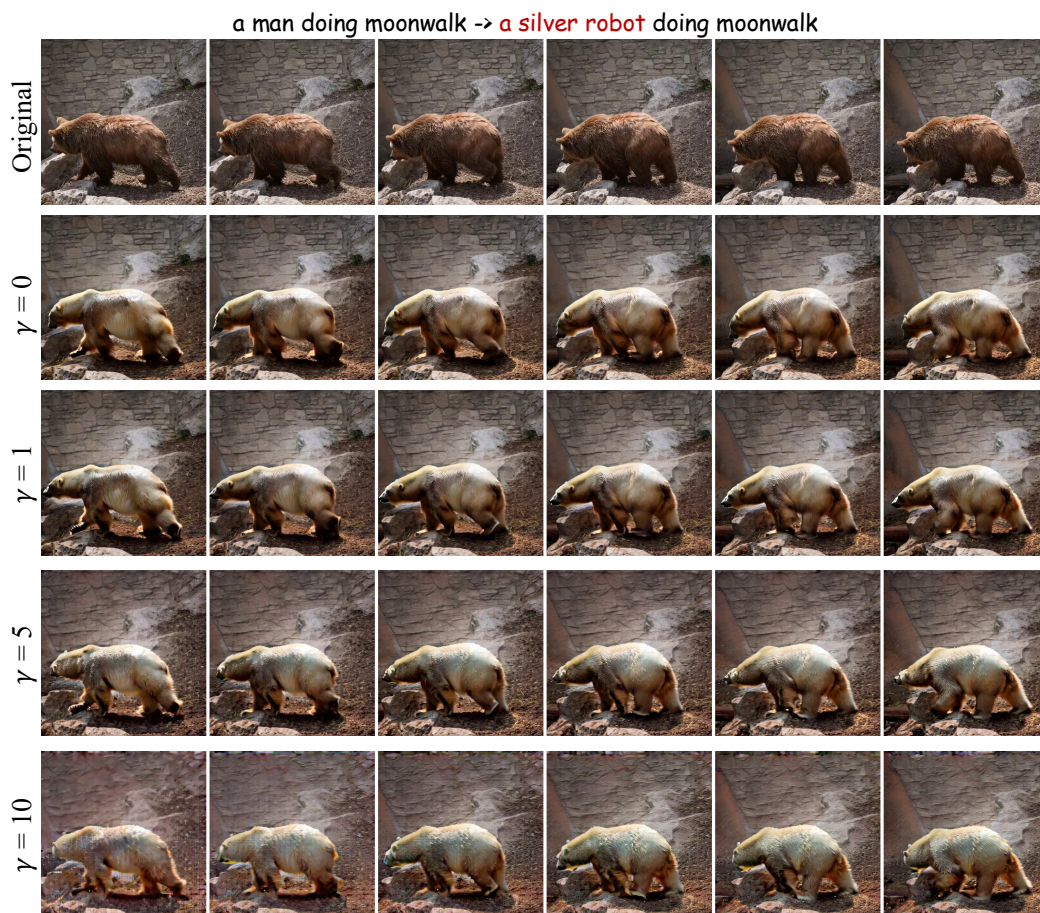
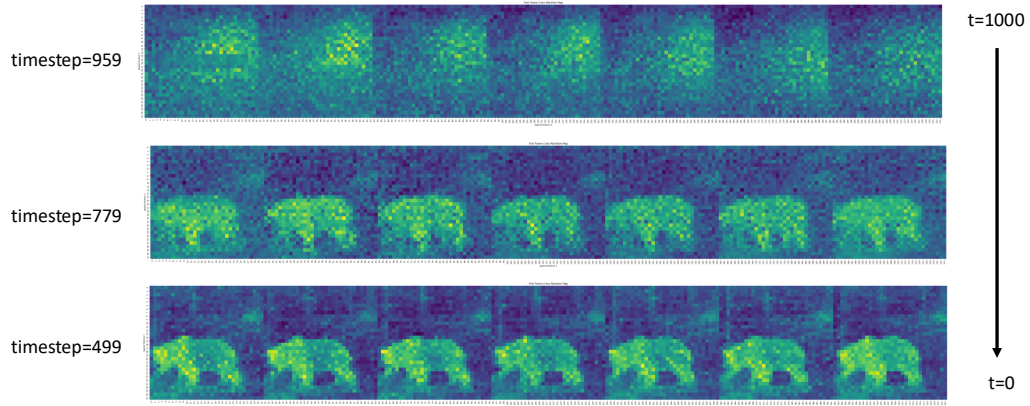
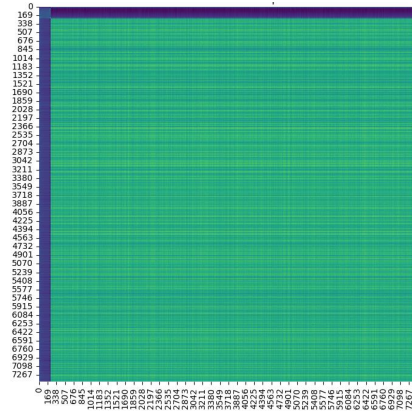


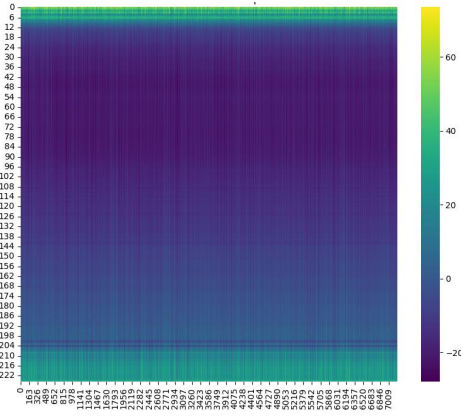
Figure F1: **Ablation results of ER on shape editing.** ER strength γ ($\gamma = 0 \rightarrow 1$ optimal, $\gamma \geq 5$ degrades).



(a) ICA Visualization



(b) FA Visualization



(c) ICA extracted from FA

Figure F2: **Implicit Cross Attention extraction and visualization.**

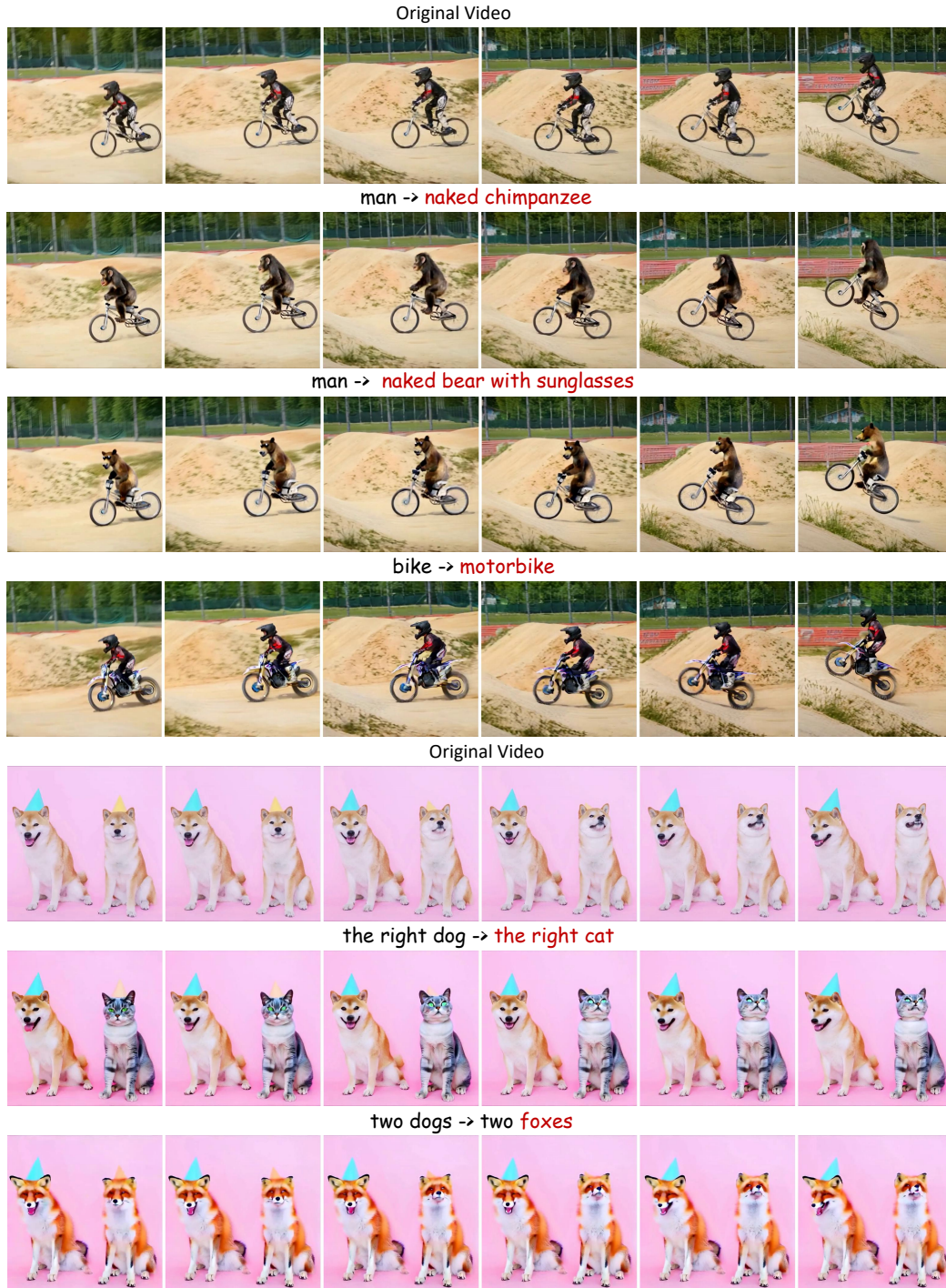


Figure F3: **Multi-object editing results.** DFVEEdit performs well on multi-object editing across various scenarios: (1) complex dynamic interactions (person-vehicle) with cluttered backgrounds, and (2) fine-grained object manipulation with detailed motions (two dogs).

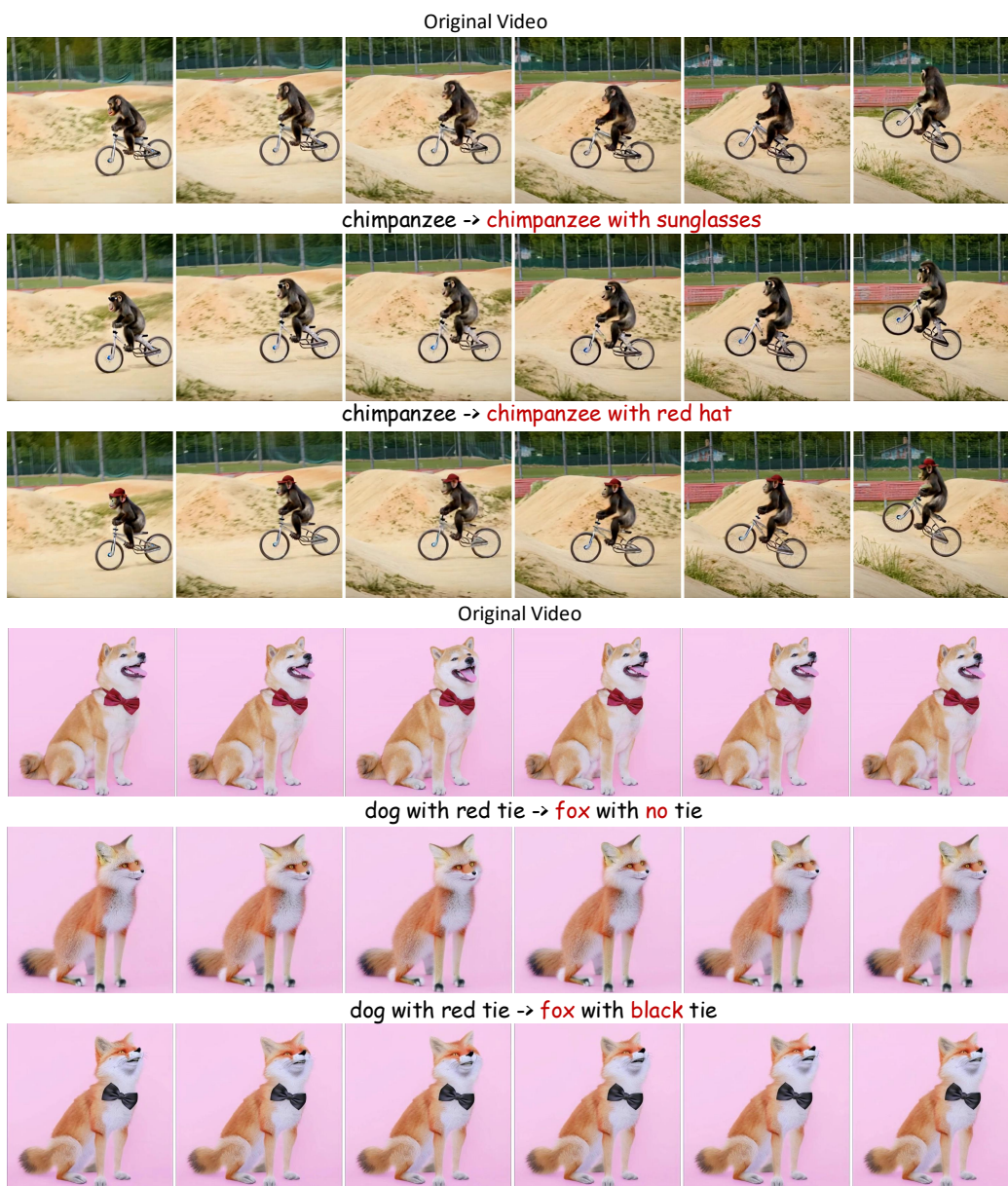


Figure F4: Attribute editing results.

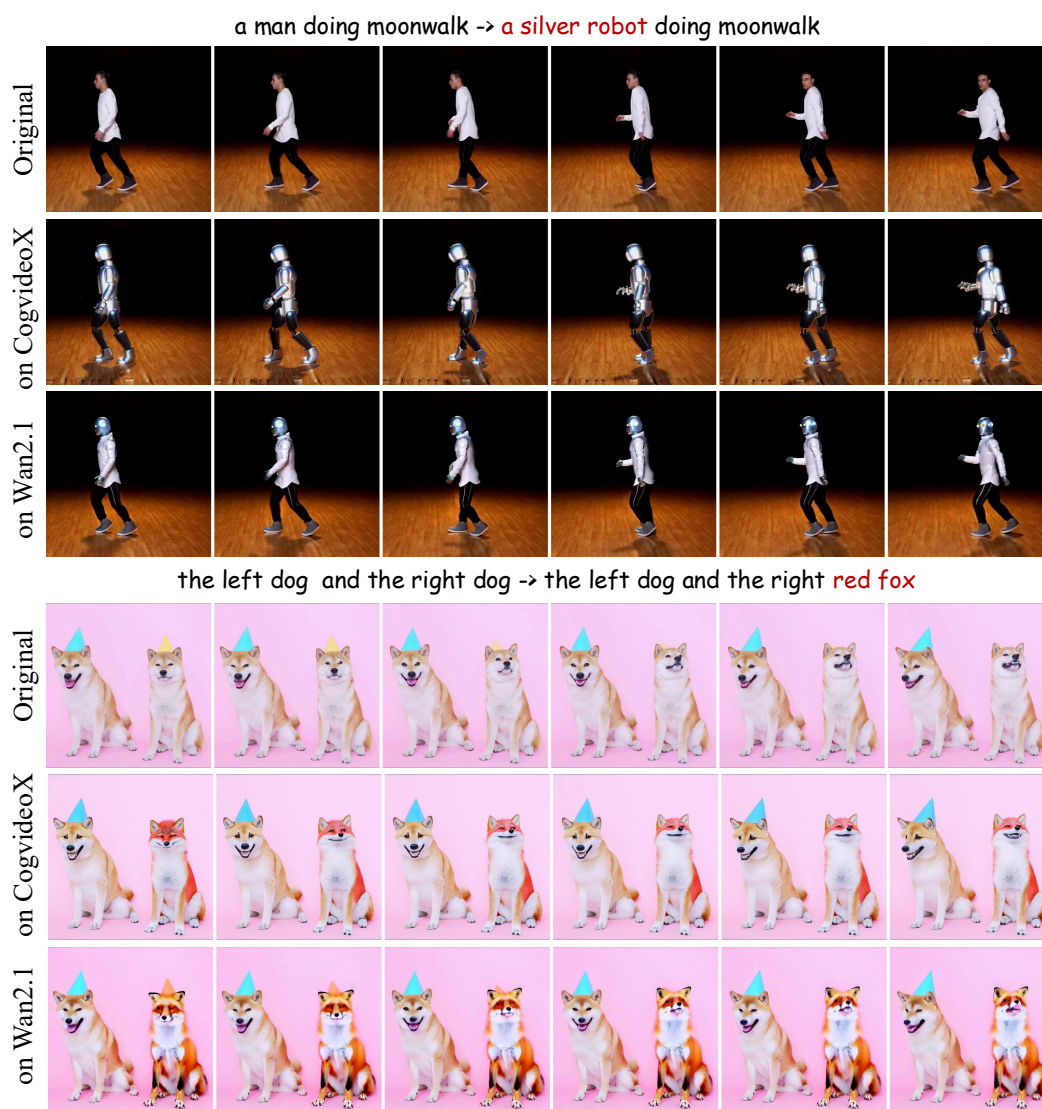


Figure F5: More extension experiment results.



Figure F6: Limitation.

References

- [1] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.
- [2] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023.
- [3] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.
- [4] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024.
- [5] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.
- [6] Pexels. Pexels free stock video clips and motion graphics. <https://www.pexels.com>. Accessed: 2025-05-15.
- [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.
- [8] 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.
- [9] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [10] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.
- [11] Lingling Cai, Kang Zhao, Hangjie Yuan, Yingya Zhang, Shiwei Zhang, and Kejie Huang. Freemask: Rethinking the importance of attention masks for zero-shot video editing. *arXiv preprint arXiv:2409.20500*, 2024.
- [12] Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. *arXiv preprint arXiv:2411.17592*, 2024.
- [13] Yuren Cong, Mengmeng Xu, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, Sen He, et al. Flatten: optical flow-guided attention for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*.
- [14] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.
- [15] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8466–8476. IEEE Computer Society, 2024.
- [16] Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. *arXiv preprint arXiv:2502.17363*, 2025.
- [17] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchu Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

- 189 [18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin
190 Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video
191 generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- 192 [19] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao,
193 Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative
194 models. *arXiv preprint arXiv:2503.20314*, 2025.