

DFVEEDIT: CONDITIONAL DELTA FLOW VECTOR FOR ZERO-SHOT VIDEO EDITING

Anonymous authors

Paper under double-blind review

ABSTRACT

The advent of Video Diffusion Transformers (Video DiTs) marks a milestone in video generation. However, directly applying existing video editing methods to Video DiTs often incurs substantial computational overhead, due to resource-intensive attention modification or finetuning. To alleviate this problem, we present DFVEEdit, an efficient zero-shot video editing method tailored for Video DiTs. DFVEEdit eliminates the need for both attention modification and fine-tuning by directly operating on clean latents via flow transformation. To be more specific, we observe that editing and sampling can be unified under the continuous flow perspective. Building upon this foundation, we propose the Conditional Delta Flow Vector (CDFV) – a theoretically unbiased estimation of DFV – and integrate Implicit Cross Attention (ICA) guidance as well as Embedding Reinforcement (ER) to further enhance editing quality. DFVEEdit excels in practical efficiency, offering at least 20x inference speed-up and 85% memory reduction on Video DiTs compared to attention-engineering-based editing methods. Extensive quantitative and qualitative experiments demonstrate that DFVEEdit can be seamlessly applied to popular Video DiTs (*e.g.*, CogVideoX and Wan2.1), attaining state-of-the-art performance on structural fidelity, spatial-temporal consistency, and editing quality.

1 INTRODUCTION

In the wave of digitization, video creation has become a dominant form of entertainment. In response, research on controllable video generation holds considerable practical importance. While Video Diffusion Transformer (DiT) models Yang et al. (2024b); Kong et al. (2024); Wang et al. (2025) have revolutionized video synthesis quality, and DiT-based image editing methods Feng et al. (2024); Kulikov et al. (2024); Zhu et al. (2025); Dalva et al. (2024); Rout et al. (2024); Jiao et al. (2025) have achieved remarkable success, video editing remains challenging in preserving spatiotemporal fidelity. Critically, existing video editing methods do not fully exploit the capabilities of Video DiTs, limiting the potential for high-quality controllable video generation.

Existing video editing techniques mainly follow two paradigms: training-based methods Singer et al. (2022); Wu et al. (2023); Shin et al. (2024); Liu et al. (2024c) and zero-shot methods Qi et al. (2023); Cai et al. (2024); Geyer et al. (2023); Zhang et al. (2023b); Yang et al.; Wang et al. (2024b). We focus on zero-shot methods for the plug-and-play flexibility and application efficiency. For training-free video editing, a high-quality pre-trained base model is crucial. Early video editing methods primarily utilized image diffusion models Rombach et al. (2022); Song et al. (2020a), which suffered from temporal inconsistencies due to the lack of capable video diffusion models. These early methods Khachatryan et al. (2023); Qi et al. (2023); Geyer et al. (2023); Yatim et al. (2024) not only had to ensure structural integrity and editing accuracy but also required significant effort to enhance temporal coherence. In contrast, methods Cai et al.

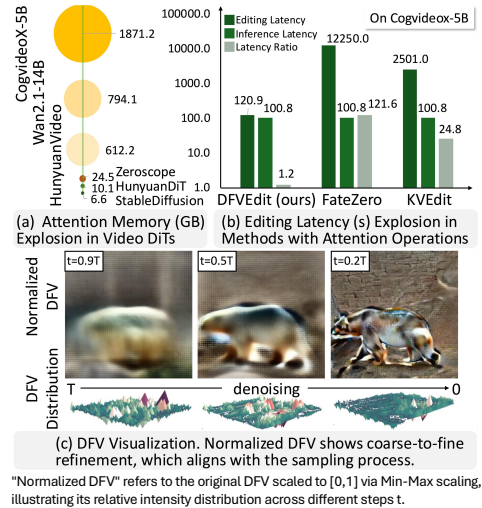


Figure 1: Key insight and motivation.

(2024); Ku et al. (2024) based on video diffusion models naturally excel in temporal consistency, leading us to leverage the latest Video DiTs Wang et al. (2025); Yang et al. (2024b); Kong et al. (2024) for video editing. Regardless of the type of base models, achieving high fidelity and temporal consistency hinges on attention engineering in most existing methods, including various attention caching and modification techniques. The key to effective attention engineering is that attentions (including keys, queries, and values) contain the spatial-temporal information of the source video, allowing for smooth editing of target regions while preserving the original content’s integrity. However, attention mechanisms now consume hundreds of gigabytes of memory (Fig. 1(a)) in Video DiTs Yang et al. (2024b); Wang et al. (2025); Kong et al. (2024), a significant increase from previous usage in Unet-based diffusion models Wang et al. (2023); Rombach et al. (2022); Song et al. (2020a) and image DiT models Li et al. (2024b); Yang et al. (2024a) at the gigabyte scale. This suggests that traditional attention engineering techniques are incompatible with Video DiTs, creating an urgent need for methods that preserve editing quality while improving computational efficiency.

Motivated by this inefficiency, we shift the focus from attention to input latents and introduce a continuous flow transformation framework, DFVEdit, for direct video latent refinement. We observe that the standard sampling process in video diffusion models—whether based on Score Matching Song et al. (2020b) or Flow Matching Lipman et al. (2022)—can be unified under a continuous flow perspective. Based on this insight, we demonstrate that editing from the source to the target video naturally forms a time-dependent flow vector field (Fig. 1(c)), which we term the Delta Flow Vector (DFV).

Building upon this foundation, we introduce the Conditional Delta Flow Vector (CDFV) to estimate the flow from source to target latent, incorporating Implicit Cross Attention Guidance (ICA) and Embedding Reinforcement (ER) to further improve editing accuracy. The CDFV in Video DiTs inherently enforces spatial-temporal dependencies while its divergence directly determines update weights. This physically grounded formulation provides two fundamental advantages over approximation-based latent-refinement approaches like DDS Hertz et al. (2023) and SDS Poole et al.: (1) *theoretical unification* by modeling both sampling and editing from the continuous flow perspective and (2) *computational efficiency* through divergence-determined and hyperparameter-free weights that eliminate heuristic scheduling and overcome low convergence issues inherent to shallow approximations. Moreover, for the seamless application to video editing, we enhanced spatiotemporal coherence by intrinsically avoiding randomness bias while incorporating ICA guidance and ER mechanisms (Fig. 5). Experiments show DFVEdit achieves at least 20× speed-up and 85% memory reduction over attention-engineering-based methods on Video DiTs (*e.g.*, CogVideoX, Wan2.1), while maintaining SOTA performance in fidelity, temporal consistency, and editing quality. Consequently, our approach offers an efficient and versatile solution for zero-shot video editing on Video DiTs.

2 RELATED WORK

Video Diffusion Transformer. Video Diffusion Transformers have evolved from early 3D-UNet-based designs Zhang et al. (2023a); Wang et al. (2023); Blattmann et al. (2023); Chen et al. (2024) to modern 3D-Transformer-based designs Peebles & Xie (2023). Advanced models such as OpenSora Zheng et al. (2024); Lin et al. (2024), CogVideoX Yang et al. (2024b), HunyuanVideo Kong et al. (2024) and Wan Wang et al. (2025) have all or part of the following key innovations: replacement of 3D-UNets with scalable 3D-Transformer blocks; integration of cross-attention and self-attention into a unified 3D-full-attention Yang et al. (2024b); Kong et al. (2024); and adoption of 3D-VAE Yang et al. (2024b) for spatiotemporal latent compression. Some Video DiTs Li et al. (2024b); Wang et al. (2025) are combined with Flow Matching Lipman et al. (2022) while others Yang et al. (2024b) adopt SDE Song et al. (2020b) samplers like DPM-solver Lu et al. (2022).

Image editing on Diffusion Transformer. With the rise of Diffusion Transformer Peebles & Xie (2023), DiT-based image editing methods Yang et al. (2024a); Li et al. (2024b) have emerged. However, directly applying image editing methods to videos often fails to address temporal consistency and motion fidelity. Additionally, adapting them to Video DiTs introduces extra challenges. Firstly, generalization limitations occur when applying methods Dalva et al. (2024); Kulikov et al. (2024); Rout et al. (2024); Jiao et al. (2025); Garibi et al. (2024); Deutch et al. (2024) that rely on rectified flow Esser et al. (2024) or distilled few-step models Sauer et al. (2024) to Video DiTs that are not combined with rectified flow or distillation techniques. Secondly, efficiency limitations are present for image editing methods Nguyen et al. (2024) that require finetuning. Furthermore, even generalized and efficient methods like DiT4Edit Feng et al. (2024) and KVEdit Zhu et al. (2025), which use

attention or key-value caching and modification, still face prohibitive computational costs due to the more massive attention overhead in Video DiTs compared to image DiTs.

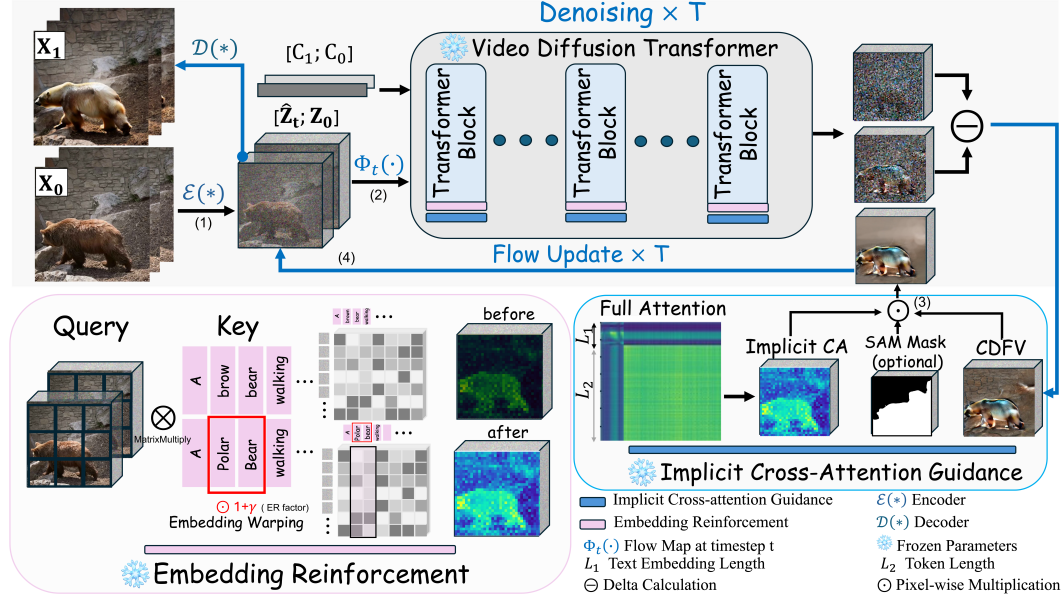


Figure 2: **DFVEdit overview.** Follow these steps for DFVEdit: (1) Encode X_0 into the latent space Z_0 , and initialize the target latent variable as $\hat{Z}_T = Z_0$. (2) Transform $[\hat{Z}_T; Z_0]$ via the flow map $\Phi_T(\cdot)$. (3) Feed the result with prompt embeddings $[C_1; C_0]$ into the Video Diffusion Transformer, compute the delta difference to obtain the CDFV at timestep T , then refine it using ER and ICA. (4) Update $\hat{Z}_T \rightarrow \hat{Z}_{T-1}$ using the enhanced CDFV, and iterate (1)-(4) until reaching $\hat{Z}_0$. (5) Decode $\hat{Z}_0$ to generate the target video X_1 .

Video editing. Video editing via diffusion models is dominated by two paradigms: training-based and training-free methods. Training-based approaches Jiang et al. (2025); Peruzzo et al. (2024); Esser et al. (2023); Gu et al. (2024); Zi et al. (2025); Wang et al. (2024a); Wu et al. (2023); Liu et al. (2024c) enhance pre-trained image diffusion models Rombach et al. (2022) with spatiotemporal modules, optimizing for complex edits but at high computational costs, limiting real-time applications. Conversely, training-free methods emphasize computational efficiency and real-time capability. Training-free video editing commonly involves two stages: latent space initialization and editing condition injection. Latent space initialization typically follows three paradigms: (1) forward diffusion with some steps for preserving low-frequency features Meng et al. (2021); Yang et al. (2023), (2) DDIM Song et al. (2020a) inversion for enabling deterministic reconstruction Qi et al. (2023); Geyer et al. (2023), or (3) direct source latent usage Hertz et al. (2023); Poole et al.. For editing condition injection, most existing zero-shot methods heavily rely on *attention engineering* to maintain spatial-temporal fidelity. For instance, FateZero Qi et al. (2023) enhances temporal consistency by caching attention maps from DDIM Song et al. (2020a) inversion and integrating them into the denoising process; TokenFlow Geyer et al. (2023) improves spatiotemporal coherence by leveraging cached attention outputs from DDIM inversion for inter-frame correspondences and incorporating extended attention blocks during denoising; VideoDirector Wang et al. (2024b) achieves fine-grained editing via SAM Kirillov et al. (2023) masks by fusing self-attention with reconstruction attention and mask guidance; and VideoGrain Yang et al. realizes complex semantic structure modifications through SAM masks while operating on complex attention map modifications. These attention-engineered methods face scalability challenges in Transformer blocks Vaswani et al. (2017), particularly for Video DiTs Kong et al. (2024); Wang et al. (2025) where attention memory demands grow dramatically (Fig. 1). Moreover, approaches Yoon et al. (2024b); Liu et al. (2024a); Ku et al.; Yatim et al. (2024) free of attention engineering suffer from structural degradation: FRAG Yoon et al. (2024b) mitigates blurring and flickering through frequency processing but compromises fidelity due to basic DDIM inversion Song et al. for source content retention; DMT Yatim et al. (2024) employs SSM Yatim et al. (2024) loss for motion transfer yet underperforms in detail preservation; and first-frame propagation

methods (e.g., StableV2V Liu et al. (2024a), AnyV2V Ku et al.) introduce accumulating artifacts without full-frame coordination. In conclusion, designing efficient and high-quality editing methods tailored for Video DiTs remains a critical challenge.

3 METHOD

Fig. 2 provides an overview of DFVEdit. Given a source video $\mathbf{X}_0 \in \mathbb{R}^{F \times 3 \times H \times W}$ comprising F RGB frames at resolution $H \times W$, together with source and target text prompts (P_0, P_1), our method supports both global stylization and local modifications (shape and attribute editing). The edited video $\mathbf{X}_1$ preserves spatiotemporal integrity in unedited regions while ensuring motion fidelity and precise alignment with P_1 . Our approach leverages two key insights: manipulating latent space is more computationally efficient than manipulating attention (Fig. 1), and editing can be modeled as the continuous flow transformation between the source and target videos (Sec 3.1). We introduce the Conditional Delta Flow Vector (CDFV) (Sec 3.2) for this transformation. To enhance video editing performance, we utilize Implicit Cross-Attention Guidance and Embedding Enforcement (Sec 3.3) to improve spatiotemporal fidelity.

3.1 UNIFIED CONTINUOUS FLOW PERSPECTIVE ON SAMPLING AND EDITING

Diffusion models include inverse and forward processes. The inverse process is typically parameterized as a Markov chain with learned Gaussian transitions, mapping noisy inputs to clean outputs. Conversely, the forward process gradually adds Gaussian noise to the clean input according to a variance schedule. As mentioned in Song & Ermon (2020; 2019); Song et al. (2020b), given a data input x , both inverse and forward processes can be regarded as overdamped Langevin Dynamics uhl (1930) (named Stochastic Differential Equation (SDE) in Score Matching Song et al. (2020b)):

$$dx_t = f(x_t, t)dt + g(x_t, t)dW \quad (1)$$

where $f(x_t, t)$ is the drift coefficient corresponds to deterministic direction and $g(x_t, t)$ is the diffusion coefficient corresponds to disturbing intensity and dW is a Wiener process and the probability density $P(x_t, t)$ can be described by introducing the Fokker-Planck equation Jordan et al. (1998) combined with the Ito's lemma Kloeden et al. (1992) and the concept of probability flow:

$$\frac{\partial P(x_t, t)}{\partial t} = -\nabla \left[\left(f(x_t, t) - \frac{g^2(x_t, t)}{2} \nabla \log P(x_t, t) \right) P(x_t, t) \right] \quad (2)$$

Eq. 2 generalizes traditional sampling methods like DDPM Ho et al. (2020) and DDIM Song et al. (2020a). ***This formulation reveals that methods based on SDE Song et al. (2020b) obey the continuity equation principle of Flow Matching Lipman et al. (2022) and can be unified under a continuous flow perspective.*** The continuous flow is characterized by a vector field $v_t(x_t) = f(x_t, t) - \frac{g^2(x_t, t)}{2} \nabla \log P(x_t, t)$, enabling state transitions from x_t to $x_{t+\Delta t}$ either through flow map Φ_t in Eq. 3 or through its Euler discretized approximation in Eq. 4:

$$\begin{cases} \frac{d}{dt} \Phi_t(x) = v_t(\Phi_t(x)) \\ \Phi_0(x) = x \end{cases} \quad (3)$$

$$x_{t+\Delta t} = x_t + \Delta t * v_t(\Phi_t(x)) \quad (4)$$

As discussed in Section 2, zero-shot video editing includes two stages: latent space initialization and editing condition injection. The first stage involves a standard sampling process. In the second stage, we derive an isomorphism with sampling process by formulating video editing as:

$$X_{t-1}^{\text{edit}} = g_{\theta_{2,t}} \left(X_t^{\text{edit}}, \underbrace{\epsilon_{\theta_1}(X_t^{\text{edit}}, t)}_{\text{Canonical Denoiser}} + \underbrace{\lambda C(X_t^{\text{edit}}, t, *)}_{\text{Control Term}} \right) \quad (5)$$

where $\{X_t^{\text{edit}}\}_{t=0}^T$ defines the state trajectory of the edited video in the sampling process; $g_{\theta_{2,t}}$ is differentiable transition function parameterized by learnable θ_2 ; ϵ_{θ_1} is pretrained diffusion model with frozen θ_1 ; $C(x, t, *)$ is the control term with intensity $\lambda \geq 0$ and optional extra input $*$. Under the Euler discretization scheme with step size $\Delta t \rightarrow 0$ and $\theta_2 = \mathcal{I}$, the discrete process in Eq. 17 converges to the controlled SDE:

$$dX_t^{\text{edit}} = \underbrace{\left[-\frac{\beta(t)}{2} X_t^{\text{edit}} + \frac{\beta(t)}{2} \nabla \log p_t(X_t^{\text{edit}}) + \lambda \frac{\beta(t)}{2} \sigma(t) C(X_t^{\text{edit}}, t, *) \right]}_{f_{\theta_1}(X_t^{\text{edit}}, t)} dt + \underbrace{\sqrt{\beta(t)}}_{g(t)} dW \quad (6)$$

where $\nabla \log p_t(X_t^{\text{edit}})$ is the score function, and $\sigma(t) = \sqrt{(1 - \alpha(t))/\alpha(t)}$ is the signal-to-noise ratio coefficient with $\alpha(t) = e^{-\int_0^t \beta(s) ds}$. **The structural isomorphism between Eq. 31 and the stochastic differential equation in Eq. 1 indicates that video editing processes can be represented within a continuous flow sampling framework**, as shown in Eq. 3 (see Appendix for more details).

3.2 CONDITIONAL DELTA FLOW VECTOR

Building upon the isomorphic correspondence between editing and sampling, we introduce the Conditional Delta Flow Vector (CDFV) to establish a direct continuous flow bridge from the source video to the target video.

Delta Flow Vector. Given the initial distribution $p(Z_T) = \mathcal{N}(Z_T; 0, I)$ for the reverse process and a clean video latent Z , Eq. 3 implies the existence of a time-dependent flow map Φ that:

$$Z = Z_T - \sum_{t=0}^T \Delta t v_t(\Phi_t(Z)) \quad (7)$$

Assuming the source and target latents $(\hat{Z}_0, Z_0)$ and their corresponding prompts (P_1, P_0) are given, we replace Z in Eq. 7 with Z_0 and $\hat{Z}_0$ respectively and define the Delta Flow Vector (DFV) as $\Delta v_t(\hat{Z}_0, Z_0) = v_t(\Phi_t(\hat{Z}_0)) - v_t(\Phi_t(Z_0))$, and the target latent $\hat{Z}_0$ can be expressed in terms of the source latent Z_0 as:

$$\hat{Z}_0 = Z_0 - \sum_{t=0}^T \Delta t \Delta v_t(\hat{Z}_0, Z_0). \quad (8)$$

Eq. 8 establishes a continuous flow directly from the source latent Z_0 to the target latent $\hat{Z}_0$, with the vector field defined as $v_t = \Delta v_t(\hat{Z}_0, Z_0)$. While prior works Han et al. (2024); Couairon et al. (2022); Hertz et al. (2023) heuristically observed that latent differences indicate editing regions, we rigorously prove this as a special case of DFV when the transformation state and vector field satisfy the continuity equation (Eq. 3).

Conditional Delta Flow Vector. The direct computation of $\Delta v_t(\hat{Z}_0, Z_0)$ is intractable since $\hat{Z}_0$ is the editing target. To resolve this problem, **we leverage the terminal condition of diffusion processes to derive an unbiased estimation of DFV**. From Eq. 2 we obtain $v_t(x_t) = f(x_t, t) - \frac{g^2(x_t, t)}{2} \nabla \log P(x_t, t)$. As t approaches T , and given that $P(x_t, t)$ is the probability density of x_t , if we set winner process of Z_0 and $\hat{Z}_0$ is equal, then $g(Z_0, t) = g(\hat{Z}_0, t)$. Consequently, as $t \rightarrow T$, both $P(Z_0, t)$ and $P(\hat{Z}_0, t)$ follow a normal distribution $\mathcal{N}(Z_T; 0, I)$ with zero mean and unit variance. Moreover, $\hat{Z}_t$ is equivalent to Z_t as $t \rightarrow T$, and we have:

$$\Delta v_t(\hat{Z}_0, Z_0) \underset{t \rightarrow T}{=} f_{\theta_1, c_1}(Z_t, t) - f_{\theta_1, c_0}(Z_t, t) \quad (9)$$

The latent $\hat{Z}_{T-\Delta t}$ can be updated using Eq. 10, which corresponds to applying the continuous flow map from $\hat{Z}_0$ as defined in Eq. 11:

$$\hat{Z}_{T-\Delta t} = Z_{T-\Delta t} - \Delta t [f_{\theta, c_1}(Z_T, t) - f_{\theta, c_0}(Z_T, t)], \quad (10)$$

$$\hat{Z}_{T-\Delta t} = \Phi_{T-\Delta t}(\hat{Z}_0). \quad (11)$$

We sequentially obtain all $v_t(\Phi(\hat{Z}_0))$ and define the Conditional Delta Flow Vector (CDFV) in Eq. 12.

$$\begin{cases} \Delta v_t(Z_0, c_0, c_1) = v_{t, c_1}(\hat{Z}_t) - v_{t, c_0}(\Phi_t(Z_0)) \\ \hat{Z}_T = \Phi_T(Z_0) \end{cases} \quad (12)$$

Theoretically, the CDFV provides an unbiased estimate of DFV. By using the CDFV as a control term, defined in Eq. 13, and integrating it into Eq. 31, we maintain a computational complexity similar to that of the basic sampling process. See the Appendix B for more details.

$$C(\hat{Z}_t, t, *) = \frac{\nabla \log P(\hat{Z}_t, t) - \nabla \log P(\Phi_t(Z_0), t)}{\sigma(t)} \quad (13)$$

3.3 SPATIOTEMPORAL ENHANCEMENT FOR CDFV

Implicit Cross-Attention Guidance. Although CDFV extracted from Video DiTs theoretically captures semantic differences between P_0 and P_1 with temporal coherence (Sec 3.2), empirical studies reveal persistent background leakage (Fig. 2). We attribute this phenomenon to the score function $\nabla_X \log p_t(X; \theta)$, which is learned by the model and may not perfectly align with theoretical expectations. This discrepancy can introduce local distributional drift in unedited regions, and such shifts have the potential to cause noticeable alterations in the background of edited videos (see Fig. 5 for examples). Segmentation masks play a crucial role in effective structure guidance, and cross-attention, as highlighted in Cai et al. (2024); Qi et al. (2023); Hertz et al. (2022), exhibit significant potential for shape editing tasks. This is attributed to their time-aware adaptability and target-following characteristics, which enhance the capability to maintain structural integrity and motion consistency over time. Although most recent Video DiTs have moved from discrete cross-attention to Full Attention Yang et al. (2024b) for more accurate spatial-temporal learning, we introduce Implicit Cross-Attention derived from Full Attention. ICA still retains the essence of traditional cross-attention and guides shape editing effectively. Given text embeddings $\mathbf{E} \in \mathbb{R}^{N \times d}$ and latent video tokens $\mathbf{B} \in \mathbb{R}^{M \times d}$, Full Attention mechanism first concatenates them to form a larger matrix $\mathbf{C} = [\mathbf{E}; \mathbf{B}] \in \mathbb{R}^{(N+M) \times d}$, each row of $\mathbf{C}$ can be considered as both Query (Q), Key (K), and Value (V). The full attention map is computed as follows:

$$\mathcal{A} = \text{Softmax} \left(\frac{\mathbf{C}\mathbf{C}^\top}{\sqrt{d}} \right) = \begin{bmatrix} \mathcal{A}_{EE} & \mathcal{A}_{EB} \\ \mathcal{A}_{BE} & \mathcal{A}_{BB} \end{bmatrix} \in \mathbb{R}^{(N+M) \times (N+M)} \quad (14)$$

We identify that the off-diagonal block $\mathcal{A}_{EB}$ or $\mathcal{A}_{BE}$ inherently encodes cross-modal interactions. Our *Implicit Cross-Attention* extracts this block of different timesteps and binarizes it into M_t . We mask $\Delta v_t(Z_0, c_0, c_1)$ with M_t to restrain the changes in the unedited region as Eq. 15. M_t can also be optionally combined with the popular SAM Kirillov et al. (2023) masks using Boolean operations.

$$\Delta v_{t, M_t}(Z_0, c_0, c_1) = M_t \odot \left[v_{t, c_1}(\hat{Z}_t) - v_{t, c_0}(\Phi_t(Z_0)) \right] \quad (15)$$

Target Embedding Reinforcement. We observe that in 3D Full-Attention, the effect of text embeddings diminishes as frame length increases. This phenomenon is particularly evident in global editing tasks such as stylization. We attribute this issue to the competition between fixed-length text tokens $\mathbf{E} \in \mathbb{R}^{N \times d}$ and an increasing number of spatiotemporal tokens $\mathbf{Z} \in \mathbb{R}^{F \times H \times W \times d}$. As the video duration grows, vectors associated with stylization embeddings become increasingly sparse across frames. This sparsity may further reduce the guidance fidelity of the text embeddings. To address these challenges, we propose Embedding Reinforcement (ER) for prompt alignment:

$$\tilde{\mathbf{E}}^{(k)} = \mathbf{E} + \gamma^{(k)} \odot \mathbf{E} \quad (16)$$

where k is used to locate the target embedding for editing, and its value is amplified by $\gamma + 1$. Specifically, we set $\gamma = 0.2$ for shape editing and $\gamma = 5$ for stylization. By reinforcing the embeddings, the cross-attention map is reweighted to focus on regions more relevant to the editing target, enhancing editing precision. Refer to the Appendix C for more method statements.

4 RESULTS

We evaluate DFVEdit on a comprehensive set of video editing tasks, including local editing (object shape and attribute modification) and global editing (style transfer) for both single (Sec. 4.1) and multiple objects editing (Appendix G). Using public videos from DAVIS2017 Pont-Tuset et al. (2017) and Pexels Pexels, our evaluation covers: (1) comparisons with state-of-the-art training-free and finetuning-based methods; (2) ablation analysis of DFVEdit’s key components; and (3) extension experiments across architectures and tasks. Additional experiment results are provided in the Appendix G, and experiment setting details are provided in the Appendix D. DFVEdit outperforms zero-shot editing methods and surpasses competing fine-tuned methods Yatim et al. (2024); Wu et al. (2023); Liu et al. (2024c) in all scenarios, excelling in structural fidelity, motion integrity, and temporal consistency. Ablation studies confirm each component’s contribution, demonstrating strong scalability across Video DiT architectures (Fig. 4, Fig. 3, Tab. T6) and generalizing well to 2D U-Net-based image diffusion models (Tab. T7, Fig. F7).



Figure 3: **Comparison.** Most methods based on attention-engineering (FateZero Qi et al. (2023), TokenFlow Geyer et al. (2023), VideoDirector Wang et al. (2024b), FreeMask Cai et al. (2024)) suffer from flickering and fail in multi-object editing. While VideoGrain Yang et al. enhances multi-object editing, it is inferior in structure consistency and motion detail fidelity (the second column). Attention-engineering-free approaches (FLATTEN Cong et al., DMT Yatim et al. (2024), ControlVideo Zhang et al. (2023b), SDEdit Meng et al. (2021)) exhibit structural infidelity. In comparison, our method achieves SOTA performance in fidelity, alignment, and temporal consistency.

Table 1: Quantitative evaluation and user study results.

Method	Consistency		Fidelity		Alignment	User Study			Computation Efficiency		
	CLIP-F $\uparrow$	E $\downarrow$ _{warp}	M.PSNR $\uparrow$	LPIPS $\downarrow$	CLIP-T $\uparrow$	Edit $\uparrow$	Quality $\uparrow$	Consistency $\uparrow$	VRAM $\downarrow$	RAM $\downarrow$	Latency $\downarrow$
SDEdit Meng et al. (2021)	0.9811	1.67	20.52	0.4090	27.46	66.57	80.45	85.66	1.01	1.13	0.87
FateZero Qi et al. (2023)	0.9289	3.09	23.39	0.2634	26.08	58.87	50.63	56.89	2.32	21.44	3.40
FreeMask Cai et al. (2024)	0.9699	2.00	29.92	0.2314	27.06	75.88	74.67	77.13	1.64	25.58	5.65
Tokenflow Geyer et al. (2023)	0.9583	1.48	29.97	0.2247	29.78	70.12	53.45	57.41	1.43	3.69	13.03
VideoDirector Wang et al. (2024b)	0.9555	2.44	28.97	0.3205	27.50	74.13	73.25	71.45	6.00	2.26	27.97
VideoGrain Yang et al.	0.9695	2.68	30.70	0.2948	27.79	76.41	79.87	70.61	2.35	2.61	13.44
FLATTEN Cong et al.	0.9510	4.89	15.91	0.3559	27.57	63.45	69.45	68.32	1.54	7.31	4.61
ControlVideo Zhang et al. (2023b)	0.9533	3.10	10.08	0.4015	27.06	56.08	55.33	59.41	8.74	1.62	9.45
DMT Yatim et al. (2024)	0.9668	3.50	15.95	0.5096	25.34	62.66	68.36	69.88	5.64	3.32	24.40
DFVEdit (on CogvideoX-5B)	0.9924	1.12	31.18	0.1886	30.84	87.65	84.56	86.98	0.95	0.86	1.20
w/o ICA	0.9922	1.25	29.33	0.1920	31.02	86.45	84.33	86.56	0.94	0.78	1.19
w/o EmbedRF	0.9913	1.13	31.15	0.1889	29.25	86.04	83.15	86.13	0.95	0.85	1.20



Figure 4: Extensive qualitative results. The extensive experiments take Wan2.1-14B Wang et al. (2025) as the base model, demonstrating the generalization of DFVEdit for Video DiTs.

4.1 COMPARISON RESULTS

Qualitative evaluation. Fig. 3 and Fig. F8 provide qualitative comparison results, showcasing our method’s superiority in structure fidelity, motion integrity, and temporal consistency over other prominent baselines. For **single object editing** (first column), FateZero Qi et al. (2023), TokenFlow Geyer et al. (2023), and VideoDirector Wang et al. (2024b) exhibit noticeable flickering, while ControlVideo Zhang et al. (2023b), FLATTEN Cong et al., and finetuning-based methods DMT Yatim et al. (2024), Tune-A-Video Wu et al. (2023), and VideoP2P Liu et al. (2024c) fail to preserve the details of unedited regions. For **multi-object editing** (second column), most methods struggle with editing accuracy; although VideoGrain Yang et al. achieves success in multi-object editing using fine-grained SAM Kirillov et al. (2023) masks, it falls short in maintaining motion detail fidelity (e.g., a mismatch between the fox and dog expressions). For **stylization** (third column), Freemask Cai et al. (2024), which is based on a UNet-based video diffusion model, performs notably well, while other methods still show inconsistencies in color tone and structural details (refer to the supplementary material for video displays). Additionally, we extended FateZero Qi et al. (2023) and KVEdit Zhu et al. (2025) directly to Cogvideo-5B Yang et al. (2024b) to compare editing quality and efficiency. Due to space limitations, please refer to the appendix for more detailed comparison results. Fig. 4 provides the extensive experiment results on Wan2.1-14B Wang et al. (2025), which also demonstrates high editing quality with respect to structure fidelity, motion integrity, and prompt alignment. Wan Wang et al. (2025) is combined with Flow Matching Lipman et al. (2022), while CogVideoX Yang et al. (2024b) is based on Score Matching Song et al. (2020b). As illustrated in both Fig. 4 and Fig. 3,

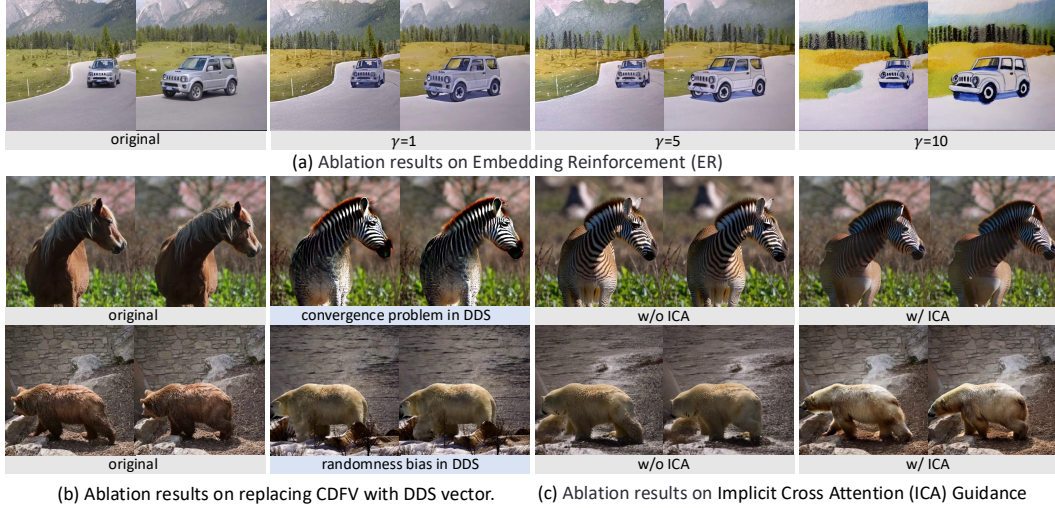


Figure 5: **Ablation.** (a)(c) demonstrate the effectiveness of ER and ICA. (b) highlights limitations of popular approximation-based latent refinement methods Hertz et al. (2023) in video editing, including: low convergence leading to unnatural changes and unpredictable convergence times; randomness bias resulting in unsatisfactory structural fidelity.

DFVEdit achieves consistent editing quality across popular Video DiTs, whether based on Score Matching Song et al. (2020b) or Flow Matching Lipman et al. (2022).

Quantitative Evaluation. In Tab. 1, we compare DFVEdit with baseline methods using both automatic metrics and user study evaluations. Detailed quantitative metrics can be found in Appendix D.2, and user study details are provided in Appendix D.3. These results highlight DFVEdit’s practical efficiency and efficacy. We extended FateZero Qi et al. (2023) and KVEdit Zhu et al. (2025) to CogVideoX-5B Yang et al. (2024b) to assess their performance. As shown in Fig. 1(b), these methods, originally designed for image diffusion, incur significant computational overhead when applied to Video DiTs. Refer to the Appendix E for more quantitative experiment results on insight. Tab. 1 shows DFVEdit’s superior performance across multiple metrics: structural consistency CLIP-F (0.9924), motion fidelity E_{warp} (1.12), unedited region fidelity M.PSNR (31.18), overall structural fidelity LPIPS (0.1886), and video-prompt alignment CLIP-T (30.84). User studies confirm its lead in editability (87.65), quality (84.56), and consistency (86.98). Additionally, it achieves notable efficiency in VRAM (0.95) and RAM (0.86) usage, with minimal latency (1.20).

4.2 ABLATION RESULTS

We evaluate the efficacy of CDFV, ICA, and ER in our ablation study. Tab. 1 reveals that omitting either the ICA or ER modules degrades performance, highlighting their indispensable roles in achieving optimal outcomes. In Fig. 5(a), we vary the Embedding Reinforcement factor γ from 1 to 10. Without reinforcement ($\gamma = 1$), stylization effects are negligible. Stylization improves as γ increases but degrades with excessively high values. Empirically, $\gamma = 5$ optimizes stylization without compromising structural fidelity or visual quality. Fig. 5(c) shows that omitting Implicit Cross-Attention Guidance leads to unintended changes in unedited regions. Incorporating cross-attention mechanisms significantly enhances structural fidelity and overall quality. In Fig. 5(b), we replace CDFV with the stochastic latent refinement vector in DDS Hertz et al. (2023). In this ablation, for the ‘horse’ experiment, ICA and ER are kept, while for the ‘bear’ experiment, they are omitted for a fair comparison. The results highlight the effectiveness of CDFV. For additional qualitative and quantitative comparison and ablation results, please refer to the Appendix F.

5 CONCLUSION

We present DFVEdit, an efficient and effective zero-shot video editing framework tailored for Video Diffusion Transformers. DFVEdit realizes video editing through the direct flow transformation of the clean source latent. We theoretically unify editing and sampling from the continuous flow perspective, propose CDFV to estimate the flow vector from the source video to the target video, and further enhance the editing quality with ICA guidance and ER mechanism. Extensive experiments demonstrate the efficacy of DFVEdit on Video DiTs.

6 STATEMENT

Ethics Statement. This work does not involve human subjects or personally identifiable information. The datasets used are publicly available and have been previously released for research purposes under appropriate licenses. Our method does not enable capabilities that are likely to cause harm when deployed, though we acknowledge that any machine learning model capable of generating realistic content could potentially be misused. No ethical approval was required for this study.

Reproducibility Statement. To ensure the reproducibility of our results, we include detailed experiment settings in the Appendix. Our code will be open-sourced upon publication. All datasets used in our experiments are publicly available. Implementation details for baseline methods are also provided to facilitate fair comparison.

REFERENCES

- On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.
- 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.
- Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.
- 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*.
- Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.
- Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow transformers. *arXiv preprint arXiv:2412.09611*, 2024.
- Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. In *European Conference on Computer Vision*, pp. 183–199. Springer, 2024.
- Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024.
- Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7346–7356, 2023.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286*, 2024.

- Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.
- Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.
- Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7621–7630, 2024.
- Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4291–4301, 2024.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). URL <https://arxiv.org/abs/2208.01626>, 3, 2022.
- Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2328–2337, 2023.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.
- Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models. *arXiv preprint arXiv:2504.13109*, 2025.
- Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.
- Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.
- Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- Peter E Kloeden, Eckhard Platen, Peter E Kloeden, and Eckhard Platen. *Stochastic differential equations*. Springer, 1992.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhui Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*.
- Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhui Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv e-prints*, pp. arXiv–2403, 2024.

- Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024.
- Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtoe: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7486–7495, 2024a.
- Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng 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*, 2024b.
- Bin Lin, Yanyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stabilizing shape consistency in video-to-video editing. *arXiv preprint arXiv:2411.11045*, 2024a.
- 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*, pp. 8599–8608, 2024b.
- 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*, pp. 8599–8608, 2024c.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.
- 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.
- Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. *arXiv preprint arXiv:2412.04301*, 2024.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
- Elia Peruzzo, Vedit Goel, Deji Xu, Xingqian Xu, Yifan Jiang, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Vase: Object-centric appearance and shape manipulation of real videos. *arXiv preprint arXiv:2401.02473*, 2024.
- Pexels. Pexels free stock video clips and motion graphics. <https://www.pexels.com>. Accessed: 2025-05-15.
- 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.
- Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*.
- 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*, pp. 15932–15942, 2023.

- 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*, pp. 10684–10695, 2022.
- Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.
- Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *Asian Conference on Machine Learning*, pp. 1215–1230. PMLR, 2024.
- Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.
- Yao Teng, Enze Xie, Yue Wu, Haoyu Han, Zhenguo Li, and Xihui Liu. Drag-a-video: Non-rigid video editing with point-based interaction. *arXiv preprint arXiv:2312.02936*, 2023.
- Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- 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.
- Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024a.
- 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*, 2024b.
- 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*, pp. 7623–7633, 2023.

- Chenglin Yang, Celong Liu, Xueqing Deng, Dongwon Kim, Xing Mei, Xiaohui Shen, and Liang-Chieh Chen. 1.58-bit flux. *arXiv preprint arXiv:2412.18653*, 2024a.
- Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.
- Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *The Thirteenth International Conference on Learning Representations*.
- 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*, 2024b.
- 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)*, pp. 8466–8476. IEEE Computer Society, 2024.
- Sunjae Yoon, Gwanhyeong Koo, Ji Woo Hong, and Chang D Yoo. Dni: Dilutional noise initialization for diffusion video editing. In *European Conference on Computer Vision*, pp. 180–195. Springer, 2024a.
- Sunjae Yoon, Gwanhyeong Koo, Geonwoo Kim, and Chang D Yoo. Frag: Frequency adapting group for diffusion video editing. *arXiv preprint arXiv:2406.06044*, 2024b.
- Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023a.
- 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*, 2023b.
- Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.
- 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.
- Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Rong Xiao, Kam-Fai Wong, and Lei Zhang. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11067–11076, 2025.

A LLM USAGE STATEMENT

We used GPT-4 (via OpenAI’s API) to assist with language polishing, grammar correction, and structural refinement of the manuscript. Specifically, the model was used to help draft and revise the introduction and related work sections, and to improve the clarity and flow of the overall writing. All prompts and generated content were carefully reviewed, critically edited, and fact-checked by the authors. The core ideas, technical content, experimental design, and final text are the responsibility of the authors and were not generated by the model.

B ADDITIONAL THEORETICAL DETAILS

B.1 REVISITING VIDEO EDITING FROM SAMPLING PERSPECTIVE

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

$$X_{t-1}^{\text{edit}} = g_{\theta_2,t} \left(X_t^{\text{edit}}, \underbrace{\epsilon_{\theta_1}(X_t^{\text{edit}}, t)}_{\text{Canonical Denoiser}} + \lambda \underbrace{C(X_t^{\text{edit}}, t, *)}_{\text{Control Term}} \right) \quad (17)$$

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

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

$$X_{t-1}^{\text{edit}} \Big|_{\substack{\lambda=0 \\ g_{\theta_2}=\mathcal{I}}} \equiv X_{t-1}^{\text{orig}} \quad (18)$$

B.2 UNIFICATION WITH VARIOUS EDITING METHODS

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

1. Inversion-based editing (like Fatezero Qi et al. (2023)):

$$g_{\theta_2,t}(a, b) = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} (a + \Delta\beta_t b) \quad (19)$$

$$C(X_t^{\text{edit}}, t, *) = \epsilon_{\theta_1}^{\text{edit}}(X_t^{\text{edit}}, t) - \epsilon_{\theta_1}(X_t^{\text{edit}}, t) \quad (20)$$

$$\Delta\beta_t = \sqrt{\frac{1 - \alpha_{t-1}}{\alpha_{t-1}}} - \sqrt{\frac{1 - \alpha_t}{\alpha_t}} \quad (21)$$

2. Latent-approximation-based editing (like DDS Hertz et al. (2023)):

$$g_{\theta_2,t}(a, b) = \text{Proj}_{\theta_2,t}(a + \eta b) \quad (22)$$

$$C(x_t, t, *) = \epsilon_{\theta_1}(x_t, t) - \epsilon_{\theta_1}(x_t, t) - \epsilon \quad (23)$$

$$\epsilon \sim \mathcal{N}(0, \sigma_t^2 I) \quad (24)$$

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 $\text{Proj}_{\theta_2,t}$ 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.

B.3 REVISITING VIDEO EDITING FROM THE CONTINUOUS FLOW TRANSFORMATION PERSPECTIVE

For the sampling process of SDE:

$$dX = \underbrace{-\frac{1}{2}\beta(t)X}_{f(X,t)} dt + \underbrace{\sqrt{\beta(t)} dW}_{g(t)} \quad (25)$$

For the inverse process of SDE:

$$dX = \left[-\frac{1}{2}\beta(t)X - \beta(t)\nabla_X \log p_t(X) \right] dt + \sqrt{\beta(t)}dW \quad (26)$$

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

$$\alpha(t) = e^{-\int_0^t \beta(s)ds}, \quad \sigma(t) = \sqrt{\frac{1 - \alpha(t)}{\alpha(t)}} \quad (27)$$

Using Taylor's expansion, we have:

$$\alpha_{t-1} \approx \alpha(t) - \dot{\alpha}(t)\Delta t \quad (28)$$

$$\frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}} \approx 1 - \frac{1}{2}\beta(t)\Delta t \quad (29)$$

$$X_{t-1}^{\text{edit}} \approx X_t^{\text{edit}} - \frac{\beta(t)}{2} (X_t^{\text{edit}} + \sigma(t)(\epsilon_{\theta_3} + \lambda C)) \Delta t \quad (30)$$

Under the Euler discretization scheme with step size $\Delta t \rightarrow 0$ and $g_{\theta_2} = \mathcal{I}$, the discrete process equation 17 converges to the controlled SDE:

$$dX_t^{\text{edit}} = \underbrace{\left[-\frac{\beta(t)}{2}X_t^{\text{edit}} + \frac{\beta(t)}{2}\nabla \log p_t(X_t^{\text{edit}}) + \lambda \frac{\beta(t)}{2}\sigma(t)C(X_t^{\text{edit}}, t) \right]}_{f_{\theta_1}(X_t^{\text{edit}}, t)} dt + \underbrace{\sqrt{\beta(t)} dW}_{g(t)} \quad (31)$$

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

$$\min_{\lambda, C} \mathbb{E} \left[\int_0^T \|C(X_t, t)\|^2 dt \right] \quad \text{s.t.} \quad dX = [f_{\theta_1}(X, t) + \lambda C(X, t)] dt + g(t)dW \quad (32)$$

$$C^*(X, t) = \frac{\nabla_X \log p_t^{\text{edit}}(X) - \nabla_X \log p_t(X)}{\sigma(t)} \quad (33)$$

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

C INNOVATIONS OVER TRADITIONAL ATTENTION-ENGINEERING APPROACHES

C.1 CONCEPTUAL AND OPERATIONAL DISTINCTIONS

- **Difference on the conceptual foundation:** We unify editing and sampling from the flow perspective, and model editing as a *continuous flow transformation* in latent space via CDFV, whereas attention engineering manipulates feature correlations via Q/K/V modifications (caching, replacing, or fusing).

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).
-

- **Difference on the operational mechanism:** CDFV performs end-to-end deformation, and ICA only infers an implicit mask from cross-attention or full-attention maps to constrain CDFV’s spatial influence, and SAM masks are also applied to CDFV optionally (used in multi-object shape editing), which means ICA and SAM masks are operated on the latent space without engaging in traditional attention engineering.
- **Difference on the architectural impact:** DFVEdit’s flow-based formulation enables *model-agnostic operation* (has been verified on 2D U-Net and various Video DiTs), while most attention engineering works fail on Video DiTs due to architectural and computational constraints.

C.2 ADVANTAGES OF ICA OVER TRADITIONAL ATTENTION ENGINEERING

ICA does not modify attention weights or feature activations but instead infers an implicit mask from attention responses to guide the CDFV. This fundamental distinction brings several key advantages:

- **Preservation of Feature Distribution:** By not altering attention outputs, ICA avoids unintended distribution shifts in unedited regions.
- **Computational Efficiency:** The mask is computed once per timestep via a forward pass, contrasting with the iterative recomputation required by attention engineering.
- **Architecture Compatibility:** As a lightweight component of CDFV, ICA can seamlessly deploy across various architectures, including 2D U-Net and Video DiTs, where traditional attention editing faces integration challenges.

C.3 USE OF SAM MASKS IN DFVEEDIT

DFVEdit uses SAM masks selectively for enhancing multi-object editing scenarios, particularly when dealing with semantic leakage issues like distinguishing between left/right objects in Fig. F10. Unlike other methods that require precise SAM masks as essential input for operation, DFVEdit operates primarily without SAM. Spatial constraints are mainly derived from ICA, producing *timestep-aware, coarse-to-fine* masks. When used, SAM masks undergo padding to soften edges and intersect with ICA maps, ensuring diffusion-aligned, soft masks that enhance editing fluidity and naturalness.

D EXPERIMENTAL SETTINGS

D.1 DATASETS

We evaluate our method on the public DAVIS2017 dataset Pont-Tuset et al. (2017) and open-source videos from Pexels Pexels, following the video editing benchmarks established by popular methods such as FateZero Qi et al. (2023) and TokenFlow Geyer et al. (2023). All experiments are conducted on one A100-80G GPU.

All experiments are conducted on 40-frame video sequences at a resolution of 512×512 . Our focus is on training-free appearance editing, encompassing both *local editing* (e.g., shape and attribute modification) and *global editing* (e.g., stylization). To comprehensively assess motion preservation capabilities, we design experiments spanning four categories of motion complexity and articulation level: 1) *High Dynamics, Low Articulation*: e.g., jeep stylization (Fig. 3) – characterized by large-scale motion with minimal structural deformation. 2) *Moderate Dynamics, Moderate Articulation*: e.g., human-to-robot translation (Fig. 3, Fig. F12) – involving intermediate limb articulation and pose transitions. 3) *Low Dynamics, High Articulation*: e.g., dogs-to-foxes translation (Fig. 3, Fig. F12, Fig. F10) – requiring preservation of fine, high-frequency non-rigid motions (e.g., ear flicks, blinking, facial expressions). 4) *High Dynamics, High Articulation*: e.g., cyclist editing (Fig. F10, Fig. F11) – combining fast motion, complex limb articulation, and micro-action transfer.

Notably, while competing methods are typically limited to editing short video clips (under 20 frames) due to memory constraints (Tab. T3), our approach demonstrates robust performance on extended 40-frame sequences without memory overflow or significant distortion. Additionally, as evidenced in Fig. F10 and Fig. 3, our method exhibits superior capability in handling complex multi-object editing scenarios. Given the inherent limitations of static figures for conveying temporal dynamics, we strongly encourage readers to consult our supplementary video. This resource provides clear demonstrations of temporal coherence, editing fidelity, and motion preservation across all evaluated scenarios.

D.2 QUANTITATIVE METRICS

We do quantitative and human evaluations with 8 quantitative metrics, including Temporal Consistency ('CLIP-F'), Warping Error (' E_{warp} ') Geyer et al. (2023), Prompt Alignment ('CLIP-T'), Masked PSNR ('M.PSNR'), Perceptual Similarity ('LPIPS'), Relative GPU Memory Consumption ('VRAM'), Relative CPU Memory Consumption ('RAM'), Relative Inference Latency ('Latency') and 3 metrics for user study, including Text Alignment ('Edit'), Overall Frame Quality ('Quality') and Temporal Consistency and Realism ('Consistency'). Specifically, **CLIP-F** calculates inter-frame cosine similarity to assess structural consistency, while E_{warp} measures warping error Geyer et al. (2023) to evaluate motion fidelity. Additionally, **M.PSNR** computes the Masked Peak Signal-to-Noise Ratio between source and target videos to gauge the fidelity of unedited regions, and **LPIPS** evaluates the Learned Perceptual Image Patch Similarity for overall structural fidelity. Moreover, **CLIP-T** quantifies the alignment between the target prompt and video through the CLIP Score Hessel et al. (2021). The results demonstrate that DFVEdit achieves superior spatial-temporal consistency, fidelity, and prompt alignment compared to other methods. Furthermore, to evaluate memory and computational efficiency, we measure Relative GPU Memory Consumption (**VRAM**), defined as the ratio of editing consumption on GPU relative to original inference consumption; Relative Inference Latency (**Latency**), which assesses the ratio of editing latency to inference latency; and Relative CPU Memory Consumption (**RAM**), measuring the ratio of editing consumption on CPU over original inference consumption.

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

Masked PSNR. We quantify structural preservation by computing Masked PSNR on unedited regions, following Liu et al. (2024b). Using 10 DAVIS Pont-Tuset et al. (2017) videos, 30 diverse prompts and corresponding segmentation annotations M provided by DAVIS, we calculate pixel-level differences between source (X_0) and edited (X_1) videos within regions identified by inverted segmentation masks $M^* = \neg M$.

$$M.PSNR(\mathbf{X}_1, \mathbf{X}_0) = PSNR(B(\mathbf{X}_1, M^*), B(\mathbf{X}_1, M^*)) \quad (34)$$

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

Relative Efficiency Metrics. We evaluate computational efficiency through three normalized metrics: Relative CPU Memory Consumption means the ratio of average CPU Memory allocated during editing to that average CPU allocated to original inference (only generation) with the same base model. Relative Inference Latency means the ratio of the latency with editing to that of the original inference latency with the same model. These relative metrics enable fair comparison across varying base model requirements. Tab. T2 reports absolute values and experimental configurations. Reported values include absolute peak allocated GPU/CPU memory consumption (MB), processing latency, frame count (F), and corresponding base models. The groups 'Stable Diffusion', 'Zeroscope', and 'CogVideoX' represent the original generation results with base models, while other groups are the editing results.

Table T2: **Absolute empirical computational efficiency results.**

Method	GPU Memory (MB)	CPU Memory (MB)	Latency (s)	F	Base Model
Stable Diffusion Rombach et al. (2022)	4134.31	2865.00	15.47	8	Stable Diffusion 1.5
Zeroscope Wang et al. (2023)	4551.32	3833.86	13.14	8	Zeroscope
CogVideoX Yang et al. (2024b)	33110.36	10522.26	100.80	41	CogVideoX-5B
SDEdit Meng et al. (2021)	33441.46	11890.15	87.69	41	CogVideoX-5B
FateZero Qi et al. (2023)	9576.13	61416.08	52.58	8	Stable Diffusion 1.5
FreeMask Cai et al. (2024)	7464.16	98059.40	74.29	8	Zeroscope
TokenFlow Geyer et al. (2023)	17040.29	159475.48	126.87	8	Stable Diffusion 1.5
VideoDirector Wang et al. (2024b)	24789.38	6475.66	432.64	8	Stable Diffusion 1.5
FLATTEN Cong et al.	6385.01	20936.98	71.35	8	Stable Diffusion 1.5
ControlVideo Zhang et al. (2023b)	36115.23	4635.36	146.19	8	Stable Diffusion 1.5
DMT Yatim et al. (2024)	42500.24	25572.34	217.54	8	Stable Diffusion 1.5
Ours	31454.84	9049.14	120.96	41	CogVideoX-5B

D.3 USER STUDY

Regarding user studies, we focus on Target Prompt Alignment (**Edit**), Overall Editing Quality including fidelity of unedited areas, minimal filtering and blurring (**Quality**), and Motion and Structural Consistency (**Consistency**). We conducted a pairwise comparison study with 20 participants evaluating 80 video-prompt pairs (30 from DAVIS, 50 from the website Pexels Pexels). Participants rated three aspects: (1) *Text Alignment* (prompt-video correspondence), (2) *Frame Quality* (visual artifacts), and (3) *Consistency* (temporal coherence and motion preservation). Scores (0-100 scale) were aggregated by trimming extremes and averaging remaining responses, yielding 1600 total ratings.

D.4 BASE MODELS

DFVEdit is focused on fully exploiting the capabilities of Video DiTs for high-quality zero-shot video editing. We take CogVideoX-5B Yang et al. (2024b) and Wanx2.1-14B Wang et al. (2025) as our base models. To the best of our knowledge, DFVEdit is the first method enabling efficient and effective zero-shot video editing on modern Video DiTs—including both score-matching (e.g., CogVideoX) and flow-matching (e.g., Wan2.1) models. Prior attention-based methods are often incompatible due to architectural and computational constraints. DFVEdit overcomes these via a lightweight, flow-field-based formulation. For a fair comparison and to show DFVEdit’s model-agnostic operation ability, we also conduct experiments on a unified 2D U-Net (Stable Diffusion 1.5) for image editing (Fig. F7, Tab. T7), demonstrating the effectiveness of DFVEdit independent of Video-DiT-specific advantages.

D.5 BASELINES

For baselines, we compare against training-free editing methods, including FateZero Qi et al. (2023), TokenFlow Geyer et al. (2023), VideoDirector Wang et al. (2024b), FreeMask Cai et al. (2024), VidToMe Li et al. (2024a) and VideoGain Yang et al., which rely on attention engineering; ControlVideo Zhang et al. (2023b), FLATTEN Cong et al., which are free of attention engineering; and

SDEdit Meng et al. (2021) (directly applied to CogVideoX-5B Yang et al. (2024b) base model for video editing), CogVideoX-V2V Yang et al. (2024b), as well as finetuning-based method DMT Yatim et al. (2024), Tune-A-Video Wu et al. (2023), Video-P2P Liu et al. (2024c).

E EXPERIMENTAL DETAILS ON INSIGHT

In Fig. 1 of the main text, we present some visualizations of our insights: (a) presents the theoretical attention memory consumption of different base models; (b) presents the theoretical inference latency of FateZero Qi et al. (2023) and KVEdit Zhu et al. (2025) when directly applying them to CogVideoX-5B, as well as the practical inference latency of DFVEdit on CogVideoX-5B. (c) reveals the evolution pattern of DFV from coarse contours to fine details, consistent with the diffusion sampling process, providing an intuitive motivation for a unified perspective on editing and generation. Here, we provide more details on these insights.

Attention memory explosion in DiT models. We analyze the memory consumption of attention mechanisms in the Unet module of diffusion models versus the Transformer module of DiT models, focusing on estimating the memory (GB) needed for storing attention score maps in float32 format per timestep. Although attention score maps are rarely computed explicitly in base models due to efficiency concerns, traditional editing methods often require their direct manipulation for attention engineering. Therefore, explicit examination of these maps helps in identifying challenges in adapting traditional editing techniques to modern DiT architectures. While our analysis centers on the memory footprint of attention score maps within a single timestep, editing methods based on attention engineering may involve caching or modifying attention maps across multiple timesteps, and we highlight the significant computation overhead when applying attention-engineering-based video editing methods to Video DiTs. As shown in Tab. T3, traditional diffusion models (e.g., Stable Diffusion and Zeroscope) exhibit multi-scale attention mechanisms with shapes varying by layer. In contrast, modern Video DiTs like CogVideoX-5B employ fixed large-scale attention $([2, 48, 11490, 11490])$, resulting in $283\times$ higher memory than SD’s maximum (7 GB vs. 1871 GB). This fundamental architectural shift explains the inefficiency of attention-based editing methods when applied to Video DiTs.

Table T3: **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 Rombach et al. (2022)*	Multi-scale	32	~ 7	float32	1
HunyuanDiT Li et al. (2024b)	[2,4096,4096]	80	~ 10	float32	1
Zeroscope Wang et al. (2023)*	Multi-scale	64	~ 25	float32	8
HunyuanVideo Kong et al. (2024)	[1,24,11520,11520]	48	~ 612	float32	41
Wanx2.1-14B Wang et al. (2025)	[1,40,11264,11264]	40	~ 794	float32	41
CogVideoX-5B Yang et al. (2024b)	[2,48,11490,11490]	40	~ 1871	float32	41

Inference Latency Comparison. We adopt a theoretical estimation approach to evaluate the inference latency for attention-engineering-based editing methods (FateZero Qi et al. (2023) and KVEdit Zhu et al. (2025)) and measure the practical inference latency of DFVEdit, since direct empirical testing of FateZero and KVEdit with Video DiT is infeasible due to GPU memory and CPU RAM constraints. First, we conduct 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.

F EXPERIMENT RESULTS ON HYPERPARAMETERS

We used the full version of DFVEdit, including both the ICA and ER components. The hyperparameters of DFVEdit are stable and easy to tune. In practice, DFVEdit introduces only a few key

hyperparameters: the fusion timesteps for ICA, the number of sampling iterations (T), and ER’s hyperparameter γ .

Seed Sensitivity and Reproducibility. The latent initialization stage is noise-free, meaning we transform a clean source video latent into a clean target latent. When computing the Conditional Delta Flow Vector (CDFV), the same Wiener process ϵ (i.e., the same random seeds) is applied to both z_t^t (target latent at t) and z_t^s (source latent) in the forward flow (Eq. 1), which leads to the cancellation of stochastic noise in the output difference (Eq. 12). Additionally, each forward pass uses a new random seed, preventing bias accumulation from fixed noise patterns. We tested DFVEdit (on CogVideoX-5B) on 20 video-prompt pairs with five different repetitive experiments to evaluate stability across random seeds. Tab. T4 reports the mean and standard deviation of key metrics, indicating low variance and consistent performance. The stability arises from our CDFV design—only the deterministic score difference drives editing, and ICA further enhances consistency by anchoring background features.

Table T4: Performance stability across 5 repetitive experiments (mean $\pm$ std).

Method	CLIP-F $\uparrow$	E_{warp} $\downarrow$	M.PSNR $\uparrow$	LPIPS $\downarrow$	CLIP-T $\uparrow$
DFVEdit	0.9924 $\pm$ 0.003	1.11 $\pm$ 0.12	31.2 $\pm$ 1.3	0.189 $\pm$ 0.012	30.8 $\pm$ 0.9

Ablation Study on Number of Diffusion Steps We thank the reviewer for this suggestion. To analyze the sensitivity of DFVEdit the total number of diffusion steps (T), we conducted an ablation study on Wan2.1-14B. As shown in Tab. T5, both editing accuracy (measured by CLIP-T $\uparrow$) and temporal consistency (E_{warp} $\downarrow$, CLIP-F $\uparrow$) stabilize when $T \geq 50$. For $T < 50$, insufficient flow field estimation leads to under-editing, manifesting as blurred or temporally unstable results. In contrast, increasing T beyond 50 yields diminishing improvements at the cost of higher computational overhead, and may result in a slight decrease in performance on background preservation. We therefore set $T = 50$ as the default, achieving an optimal trade-off between efficiency and quality. This value is also consistent with common practice in high-fidelity video generation, where moderate step counts suffice under advanced solvers.

Table T5: Ablation study on the number of diffusion steps T .

Total Timesteps T	CLIP-F $\uparrow$	E_{warp} $\downarrow$	M.PSNR $\uparrow$	LPIPS $\downarrow$	CLIP-T $\uparrow$
10	0.9731	2.1305	30.56	0.1793	29.91
20	0.9833	2.1100	30.61	0.1688	30.66
50 (default)	0.9950	1.0600	31.23	0.1568	31.34
100	0.9954	1.4175	30.96	0.1569	31.32
150	0.9973	1.3881	30.73	0.1571	31.36
200	0.9953	1.3672	30.66	0.1578	31.50

Embedding Reinforcement. Due to space limitations, 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. F9, 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 ICA masks. 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 Cai et al. (2024), with the layer index selection method also following FreeMask), and (2) SAM masks with edge padding. Fig. F6 shows an example of ICA extraction and visualization. This strategy operates in two phases: ICA guidance during early

Table T6: Quantitative comparison with additional state-of-the-art video editing methods.

Method	CLIP-F $\uparrow$	E_warp $\downarrow$	M.PSNR $\uparrow$	LPIPS $\downarrow$	CLIP-T $\uparrow$
VideoP2P Liu et al. (2024c)	0.9624	3.28	17.38	0.4531	27.26
Tune-A-Video Wu et al. (2023)	0.9612	3.35	16.67	0.4545	27.14
CogvideoX-V2V Yang et al. (2024b)	0.9812	1.67	20.53	0.4092	27.46
VidToMe Li et al. (2024a)	0.9737	2.96	22.52	0.3062	27.32
DFVEdit (CogvideoX-5B)	0.9924	1.12	31.18	0.1886	30.84
DFVEdit (on Wan2.1-14B)	0.9950	1.06	31.23	0.1568	31.34

denoising ($t = T \rightarrow 0.4T$) preserves shape flexibility while reducing background leakage, followed by SAM-based mask guidance ($t = 0.3T \rightarrow 0$).

G ADDITIONAL EXPERIMENTAL RESULTS

Extended Comparison Experiments. To ensure a comprehensive evaluation, we additionally compare with CogVideoX-V2V Yang et al. (2024b) on the Video DiT backbone (CogVideoX-5B), state-of-the-art fine-tuning-based editing methods including Tune-A-Video Wu et al. (2023), VideoP2P Liu et al. (2024c), and the latest competing method VidToMe Li et al. (2024a). As illustrated in Tab. T6 and Fig. F8, DFVEdit demonstrates superior performance. Furthermore, qualitative results using Wan2.1-14B are provided in Figs. 4 and F12, with corresponding quantitative metrics in Tab. T6. These results confirm that DFVEdit achieves enhanced outcomes when leveraging this advanced Video DiT backbone.

Image Editing Experiments We conducted experiments on PIE-Bench Ju et al. (2023) using Stable Diffusion 1.5, comparing DFVEdit with Instruct-pix2pix Brooks et al. (2023), PnP Tumanyan et al. (2023), and P2P Hertz et al. (2022) (PnP and P2P have adopted DirectInversion Ju et al. (2023) techniques). We evaluated DFVEdit and T2I baselines on two representative subsets: *1_change_object_80* and *9_change_style_80*, using standard metrics—Distance (overall structure coherence), PSNR/LPIPS/SSIM (background preservation), and CLIP Similarity (whole alignment)—following DirectInversion. Results in Tab. T7 and Fig. F7 show the competitive performance and practical cross-modal generalization ability of DFVEdit. Our core contribution remains enabling efficient, zero-shot editing on modern Video DiT models, where DFVEdit overcomes the computational and architectural limitations of attention-based methods without requiring fine-tuning.

Table T7: Quantitative results on PIE-Bench Ju et al. (2023) for image editing.

Method	Structure		Background		
	Distance $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	CLIP Similarity $\uparrow$
DFVEdit	0.0167	23.24	0.1225	0.6935	29.62
Instruct-pix2pix Brooks et al. (2023)	0.0258	21.56	0.1308	0.6321	26.13
P2P Hertz et al. (2022)	0.0186	22.38	0.1257	0.6416	29.07
PnP Tumanyan et al. (2023)	0.0279	18.06	0.1353	0.6135	29.02

On multi-objects editing. As shown in Fig. F10, 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.

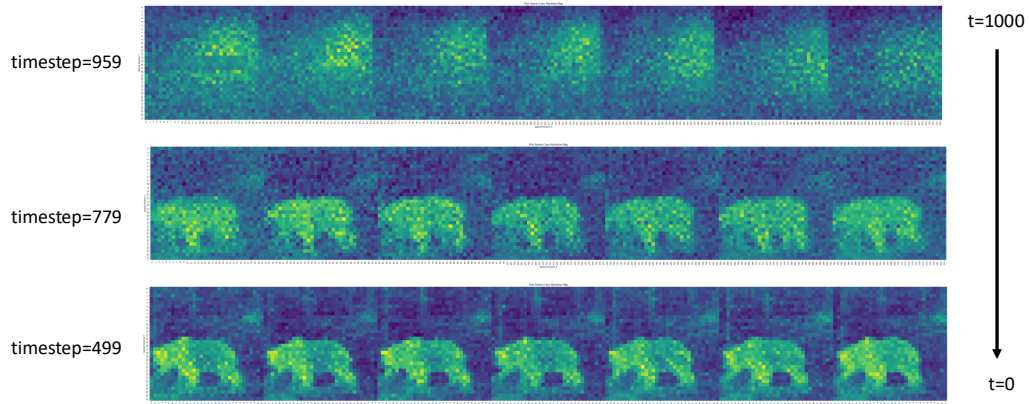
More results on attribute editing. Due to space limitations, we include the visualization results of attribute editing in the Appendix. As shown in Fig. F11, our method demonstrates satisfactory performance on attribute editing, enabling the natural and seamless integration of both added and removed 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, Fig. F12 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.*

H LIMITATIONS AND FUTURE WORK

As illustrated in Fig. F13, our method exhibits several limitations common to zero-shot video editing. First, maintaining perfect detail fidelity in non-edited regions remains challenging. While our approach demonstrates superior fidelity preservation compared to existing methods, some detail loss persists even with ICA guidance. This issue is partially attributable to VAE compression artifacts, a fundamental bottleneck in latent diffusion models that particularly affects high-frequency textures and fine structures. In addition, our method has limited capability for large shape variations (e.g., bicycle-to-car transformations). The flow-based formulation and the underlying DiT's structural dependencies restrict applications to shape editing tasks with minor layout changes. Generating entirely new geometries while preserving background consistency and motion dynamics remains an open challenge in video editing. Furthermore, DFVEdit is primarily designed for appearance editing and currently offers limited support for non-rigid shape modifications Yoon et al. (2024a) or drag-based interactive editing Teng et al. (2023); Deng et al. (2024). Future work could explore incorporating explicit structural guidance (e.g., depth, optical flow) or layered compositing strategies to address these limitations, potentially enabling more extreme shape transformations and interactive editing capabilities.



(a) ICA Visualization

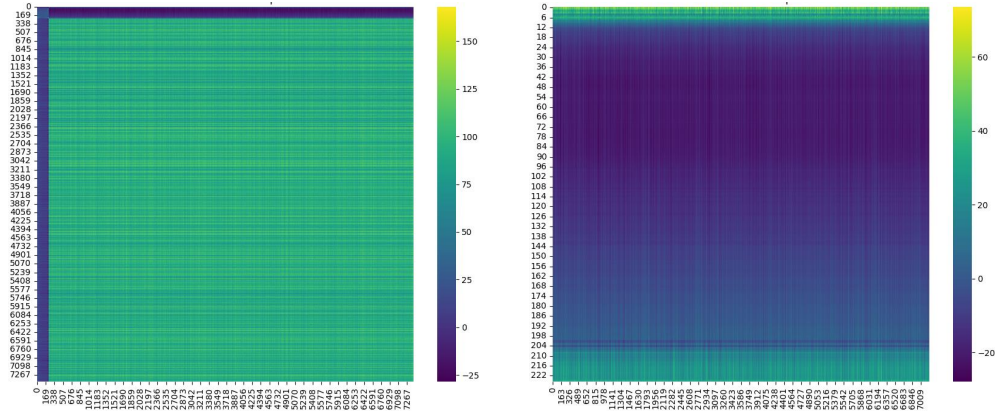


Figure F6: Implicit Cross Attention extraction and visualization.



Figure F7: Qualitative comparison results on image editing with PIE-bench.

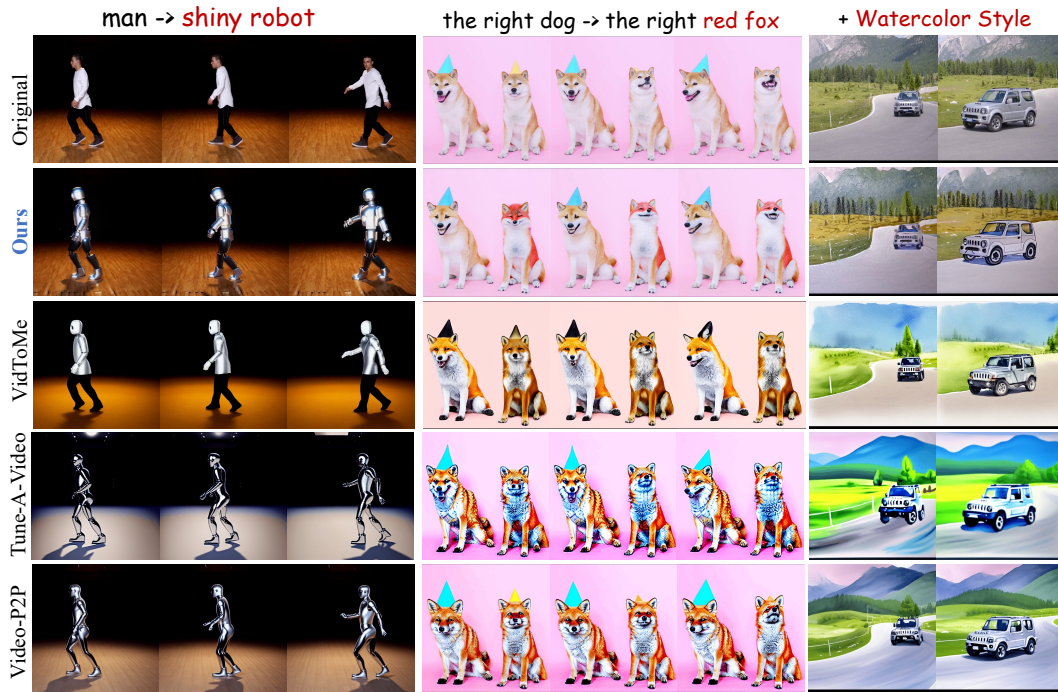


Figure F8: More comparison results.

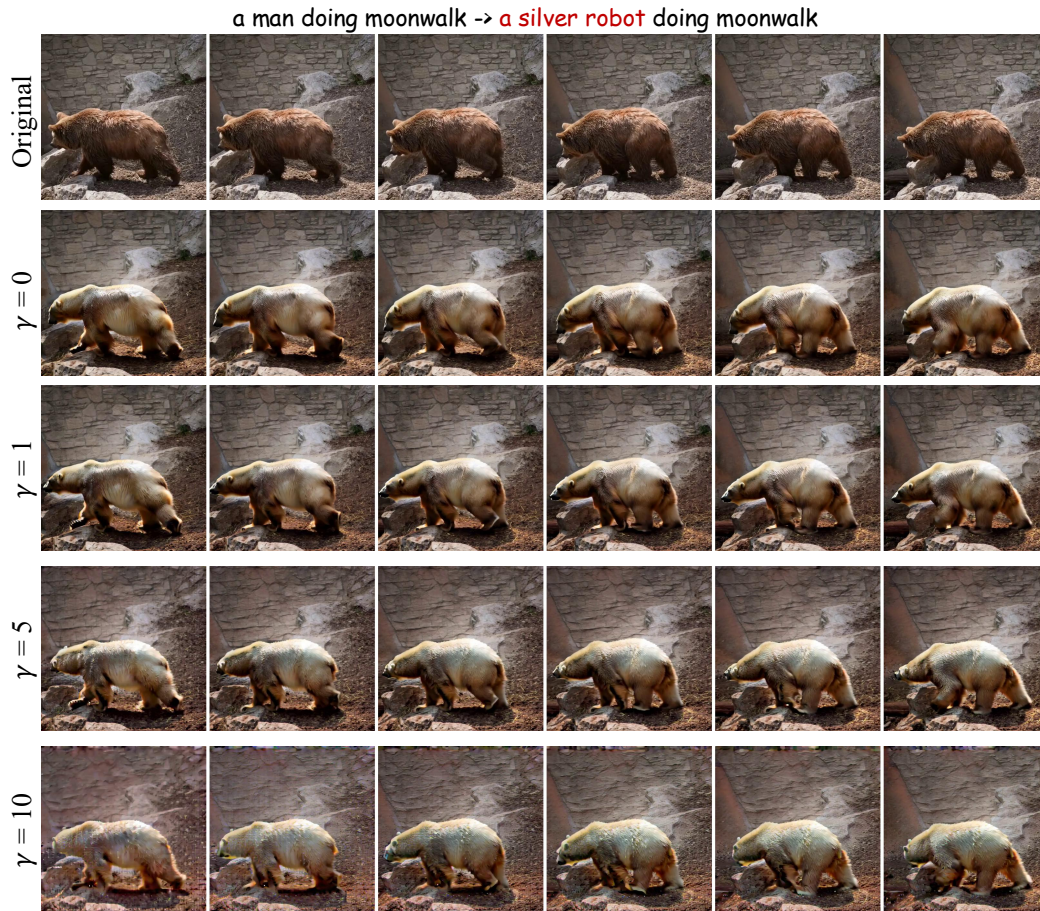


Figure F9: **Ablation results of ER on shape editing.** ER strength γ ($\gamma = 0 \rightarrow 1$ optimal, $\gamma \geq 5$ degrades).

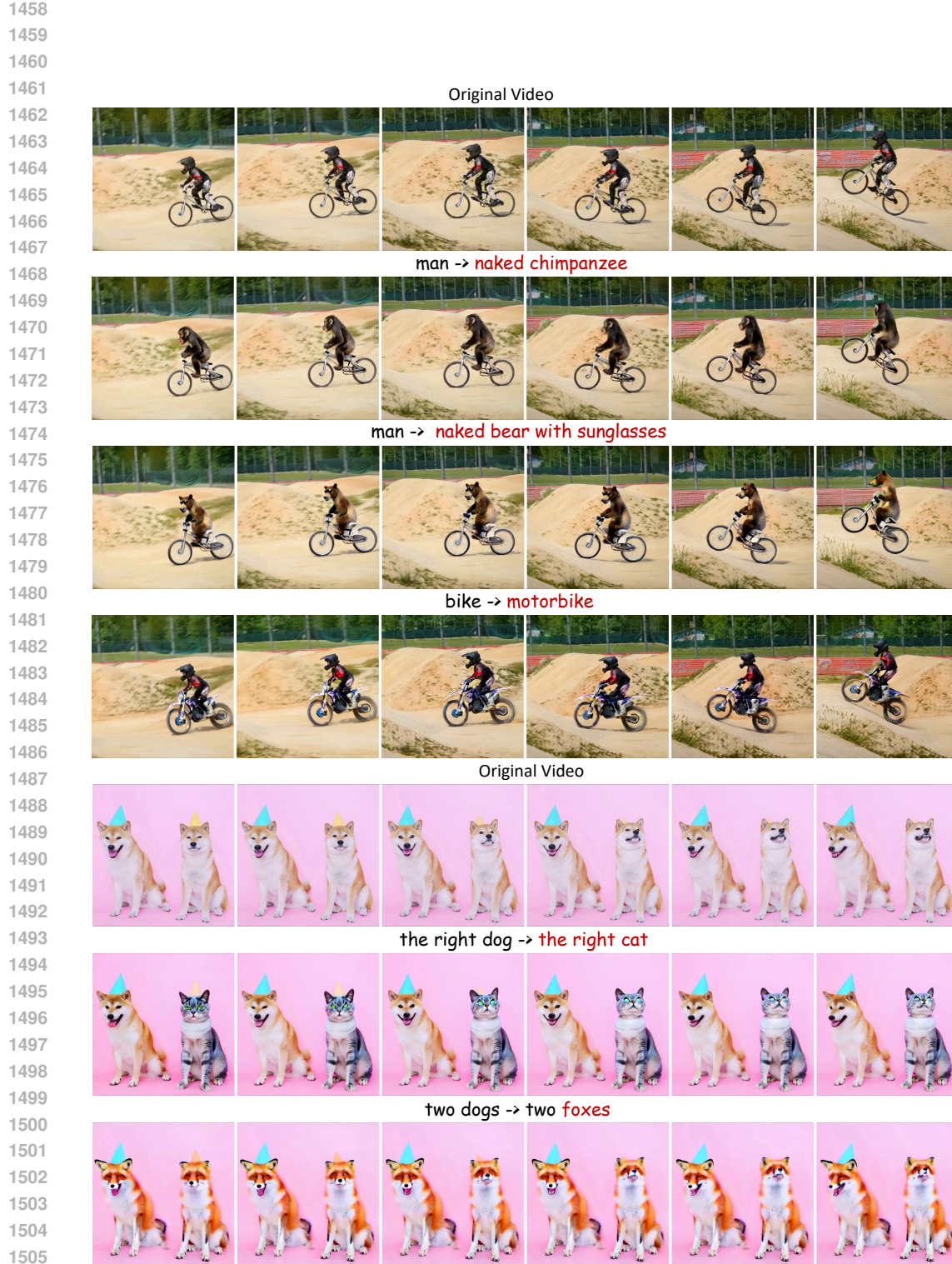


Figure F10: **Multi-object editing results.** DFVEdit performs well on multi-object editing across various scenarios: (1) complex dynamic interactions (person-vehicle) with cluttered backgrounds, and (2) fine-grained object manipulation with detailed motions (two dogs).

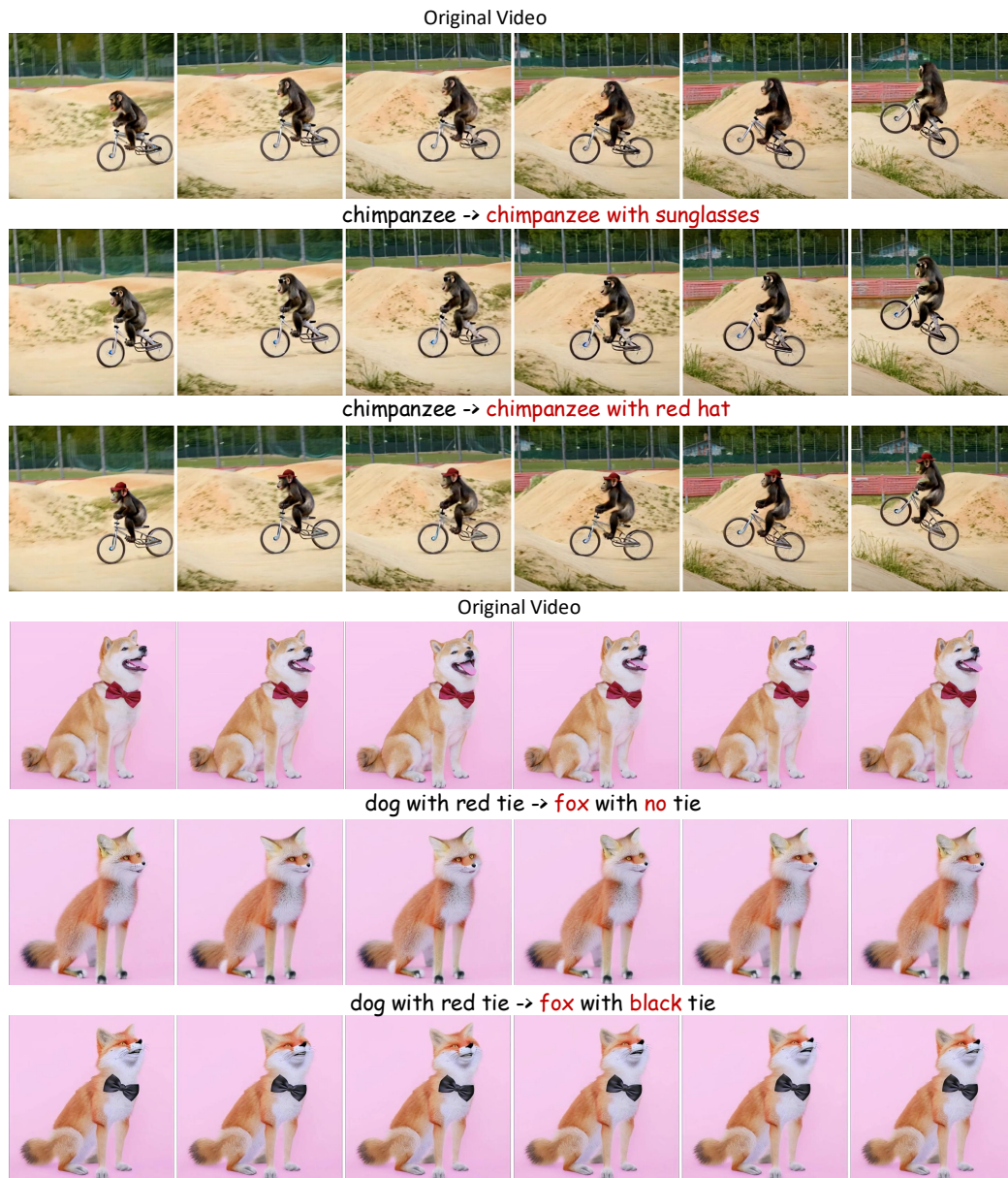


Figure F11: Attribute editing results.



Figure F12: More extension experiment results.



Figure F13: Limitation.