

WAVE: Warping DDIM Inversion Features for Zero-shot Text-to-Video Editing

Yutang Feng^{1*}, Sicheng Gao^{1,3*}, Yuxiang Bao^{1*}, Xiaodi Wang², Shumin Han^{1,2†}, Juan Zhang^{1†}, Baochang Zhang^{1,4}, and Angela Yao³

¹ Beihang University, Beijing, China

² Baidu VIS

³ National University of Singapore

⁴ Zhongguancun Laboratory, Beijing, China

{yutangfeng,zhang_juan}@buaa.edu.cn, hanshumin@baidu.com

Abstract. Text-driven video editing has emerged as a prominent application based on the breakthroughs of image diffusion models. Existing state-of-the-art methods focus on zero-shot frameworks due to limited training data and computing resources. To preserve structure consistency, previous frameworks usually employ Denoising Diffusion Implicit Model (DDIM) inversion to provide inverted noise latents as guidance. The key challenge lies in limiting errors caused by the randomness and inaccuracy in each step of the naïve DDIM inversion process, which can lead to temporal inconsistency in video editing tasks. Our observation indicates that incorporating temporal keyframe information can alleviate the accumulated error during inversion. In this paper, we propose an effective warping strategy in the feature domain to obtain high-quality DDIM inverted noise latents. Specifically, we shuffle the editing frames randomly in each timestep and use optical flow extracted from the source video to propagate the latent features of the first keyframe to subsequent keyframes. Moreover, we develop a comprehensive zero-shot framework that adapts to this strategy in both the inversion and denoising processes, thereby facilitating the generation of consistent edited videos. We compare our method with state-of-the-art text-driven editing methods on various real-world videos with different forms of motion. The project page is available at <https://ree1s.github.io/wave/>.

Keywords: Text to video editing · DDIM inversion · Flow-guided warping

1 Introduction

Recent text-driven image diffusion models such as DALL·E [33], Stable Diffusion [35], Imagen [14] can generate high-quality and realistic images. With these powerful models, diverse image editing techniques have been developed *e.g.* colorization [36], inpainting [1], and image-to-image translation [41], all with the

*Equal contribution. Work done when Y. Feng was an intern at Baidu.

†Corresponding authors.



Fig. 1: WAVE, a novel zero-shot text-driven video editing framework, excels in producing high-quality and temporal consistent edited videos by leveraging optical flow extracted from the source videos. Given a source video and a target textual prompt, WAVE employs a flow-guided feature warping strategy during both DDIM inversion and joint editing processes for powerful video editing.

convenience of textual descriptions. Considering the mainstream status of short videos in social media and online platforms, there is a similar demand for text-based video editing.

Initially, researchers attempted to replicate the training paradigm of image generation with text-video datasets [14, 16, 38, 47]. However, the computational expense coupled with the limited dataset size makes a direct translation difficult to achieve. The more common paradigm is to utilize image models for individual videos in a zero-shot manner without training or fine-tuning on video data. Current zero-shot methods [4, 11, 44] adopt a three-phase editing process. First, they invert a sequence of frames from the video with a source text prompt into the domain of pretrained image generative models. Afterward, they conduct temporal consistent keyframe sampling with inter-frame extended attention. The final step is to propagate in the feature domain to arrive at the edited video [11].

Inspired by these methods, we aim to design a zero-shot video editing method using off-the-shelf image diffusion models. The inversion process benefits the video editing result by providing a sequence of temporally coherent initial latents. These are more likely to be recovered into edited sequences with high fidelity and visual consistency. However, as illustrated in Fig. 2, the direct inversion process may lead to an undesired gap between the ideal and actual inverted latents, due to the potential error accumulation.

To tackle this problem, previous approaches focus on enhancing semantic consistency across frames [22] or fusing the attention maps in the inversion process [31]. Even so, as shown in Fig. 2, the error of inverted latents still accumulates over each timestep. The unaddressed error accumulation carries over

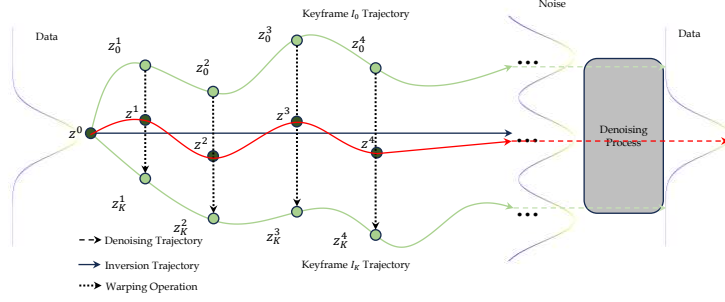


Fig. 2: Our proposed DDIM Inversion overview. The black and green lines represent the ideal versus possible naive DDIM inversion trajectories, respectively. Our proposed method, depicted by the red line, aims to minimize the deviation from the ideal trajectory by leveraging inter-frame information with a flow-guided warping strategy. This approach ensures high-quality reconstruction, as evidenced by the convergence of our trajectory to the target data manifold following the same denoising process.

into the subsequent steps. This oversight leads to the manifestation of unfaithful spatial textures and brings a significant flickering phenomenon in video editing tasks.

This work targets limiting the error accumulation over the inversion timesteps. To that end, we propose a simple yet effective zero-shot approach and incorporate a flow-guided warping strategy in the feature domain during inversion. The motivation behind our method is that the accumulated error could be significantly alleviated by utilizing inter-keyframe information. Specifically, at each time step, we randomly shuffle the editing keyframes, and warp the latent feature from the first frame to the others with optical flow extracted from the source video. In this manner, the potential error accumulating in the inversion process is mitigated. This improvement lays high-quality structure guidance for the subsequent stages of editing and propagation, enhancing the overall quality and consistency of the video editing pipelines. Moreover, the feature warping strategy is still adopted in the denoising process, in the sense that a symmetric and unified framework could bring advantages to the coherence of edited video.

Our contributions are summarized as follows:

- We introduce a flow-guided warping strategy, specifically designed to efficiently propagate latent features across randomly shuffled frames to reduce the error of accumulation of inverted noise latents.
- We propose a unified framework that applies this strategy in both the inversion and denoising stages, to benefit video editing consistency.
- Our method has shown remarkable editing results across extensive experiments; compared to the previous zero-shot methods our edited videos are temporally coherent and flicker-free.

2 Related Work

Text-driven Generation. In past years text-to-image generation models were based on Generative Adversarial Networks (GANs) [21,32,34,51,57]. More recent works adopt diffusion models [15] for various generation tasks [10,19,23,28,37,48–50]. Some such as DALL-E-2 [33] and Imagen [14] complete the denoising process directly in the pixel space, while Latent Diffusion Models (LDMs) [35] implement an auto-encoder to improve the representation efficiency in the low-dimensional latent space. Based on LDMs, we employ a pretrained image diffusion model for our inverting and editing stages.

Text-driven video generation attempts to adopt and extend these powerful image generative models into the spatio-temporal domain [16,38,55]. LDMs [3] propose a temporally consistent manner to align images by training extra temporal layers. PVDm [47] adopts a triplane representation for encoding videos and project videos as 2D latents for training. However, training video generation models requires large-scale video data and exorbitant computing resources.

Text-driven Editing Based on the above described generation models, text-driven editing is a popular research area and many works [1,7,8,20,52] adopt these generative priors to edit images or videos with given text prompts. To utilize unconstrained text-to-image models for controlled content creation without any training or fine-tuning, SDEit [25] enables the editing of user-provided images using free text prompts. PnP-Diffusion [41] also avoids involving any training or fine-tuning by manipulating spatial features and the corresponding self-attention directly. Meanwhile, the method of manipulating features motivates us to achieve coherence among keyframes within the feature domain. While these image editing methods achieve better-edited results, the challenges posed by inter-frame flickering and inconsistency render their application problematic.

For text-to-video editing, a surge of works demonstrate how to train spatio-temporal diffusion models on large-scale datasets [3,9,27,38]. However, recent approaches are primarily focused on bypassing the extensively time-consuming training phase of traditional models. For instance, Tune-A-Video [44] just fine-tunes the inflated cross-frame attention on a single video to learn continuous motion to facilitate new video editing. As a further work, FateZero [31] introduces a zero-shot approach via an attention blending block to enable shape-aware editing based on pretrained Tune-A-Video models.

While [44] and [31] maintain the structure of a single frame, they often fall short of controlling visual consistency across the entire video. Utilizing ControlNet [52], Text2Video-Zero [18] and ControlVideo [54] preserve the content structure but lack temporal consistency. To tackle this challenge, previous works [2,6,11] have demonstrated that the inter-frame correspondences extracted from the source video can be utilized to enhance motion consistency in video editing. In this work, we further unify the DDIM inversion and joint editing processes to ensure more coherent video editing.

DDIM Inversion DDIM inversion is the process of performing DDIM sampling in reverse order. In Generative Adversarial Networks (GANs), the inversion finds a latent code that corresponds to a given image [12, 29, 45, 56] and is extended in many image tasks as a powerful prior [5, 43]. For diffusion models, however, the inversion process needs an appropriate noise map and a conditional vector for a generated image. For instance, Null-text inversion [26] aligns the trajectory of diffusion latents with denoising latents. This inversion is frequently employed in video editing tasks as well [11, 44]. Nevertheless, directly adding Gaussian noise and denoising may lead to substantial alterations in the image textures over the iterative process, thereby exacerbating the temporal inconsistency. Inspired by [26], [22] optimizes a shared unconditional embedding for all frames to improve the video inversion quality, while the tuning strategy may break the pretrained generative ability. FateZero [31] does not require extra tuning by fusing the attention maps in the inversion process. However, it does not address the issue of the accumulated errors that arise during this process. In contrast, our flow-guided warping strategy can eliminate error accumulation by calculating the average.

3 Preliminaries

3.1 Diffusion Models

Stable diffusion [35] represents a significant advancement in diffusion models that function within the latent space. It uses a pretrained autoencoder to encode images into latent space and reconstruct them into high-resolution images. The process initiates with an initial latent signal z_0 , corresponding to the latent of an input image I_0 . This procedure iteratively advances through a sequence of defined steps, with the forward diffusion process of z_t from z_{t-1} defined as follows:

$$q(z_t | z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t} z_{t-1}, \beta_t I), \quad (1)$$

where q is the forward diffusion process and β_t represents the variance schedule at each timestep t . The total number of timesteps T denotes the extent of the training duration for the diffusion model. The reverse process of z_{t-1} from z_t uses a conditional denoising model U-Net ϵ_θ defined as:

$$p_\theta(z_{t-1} | z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, \tau, t), \Sigma_\theta(z_t, \tau, t)), \quad (2)$$

where p indicates the inference process in terms of Gaussian conditional distributions and τ denotes encoding of the text prompt and μ_θ and Σ_θ are Gaussian parameters estimated by the denoising model ϵ_θ . In Stable Diffusion, the denoising model ϵ_θ containing cross-attention and self-attention modules [42] is trained to predict the noise with the following objective function:

$$\min_{\theta} E_{z_0, \varepsilon \sim N(0, I), t \sim \mathcal{U}(1, T)} \|\varepsilon - \varepsilon_\theta(z_t, t, \tau)\|_2^2, \quad (3)$$

where, $\mathcal{U}$ represents a uniform distribution and z_t is a noisy latent of z_0 at timestep t .

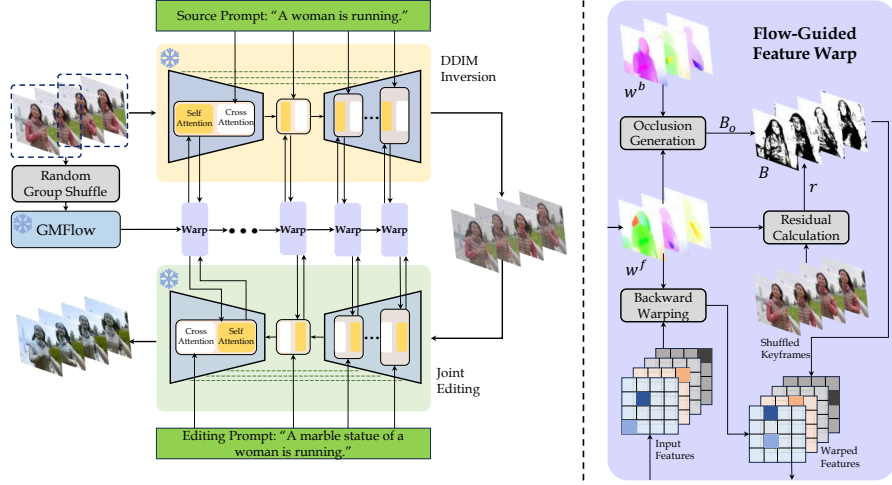


Fig. 3: Our pipeline. We demonstrate the overall framework of **WAVE** in the left part of the figure. We warp the input of self-attention modules and residual blocks between randomly shuffled keyframes. In the right part, we show the workflow of our flow-guided feature warp strategy, including binary mask generation, occlusion generation, and feature alignments.

3.2 DDIM Inversion

At inference, existing methods usually seek structure guidance from the source video through DDIM inversion [39]. A deterministic latent representation z_0 can be converted from a random Gaussian noise during DDIM sampling in a sequence of timestep $t : T \rightarrow 1$:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \frac{z_t - \sqrt{1 - \alpha_t} \varepsilon_\theta}{\sqrt{\alpha_t}} + \sqrt{1 - \alpha_{t-1}} \varepsilon_\theta, \quad (4)$$

where α_t denotes $\prod_{i=1}^t (1 - \beta_i)$. The DDIM inversion technique is designed to map a clean latent z_0 onto its corresponding noised latent $\hat{z}_T$. This mapping occurs through a sequence of reversed timesteps $t : 1 \rightarrow T$:

$$\hat{z}_t = \sqrt{\alpha_t} \frac{\hat{z}_{t-1} - \sqrt{1 - \alpha_{t-1}} \varepsilon_\theta}{\sqrt{\alpha_{t-1}}} + \sqrt{1 - \alpha_t} \varepsilon_\theta, \quad (5)$$

which can serve as a structural guidance for the requirements of reconstruction and translation processes.

3.3 Optical Flow Estimation

Consider two consecutive RGB frames of size $H \times W$ from the source video with their 2D coordinates of pixel grid G . We use a pre-trained optical flow

estimation network [46] to estimate the optical flow V between the frames. The flow estimation takes on a global matching formulation to address the large displacement issue in video translation tasks which is suitable for our subsequent random shuffle warping strategy. Specifically, for feature maps of the two frames, each pixel is compared based on their feature similarity to estimate a matching distribution $M \in \mathbb{R}^{H \times W \times H \times W}$ with the softmax operation. The optical flow V can be obtained by computing the difference between the corresponding pixel coordinates:

$$V = MG - G \in \mathbb{R}^{H \times W \times 2}, \quad (6)$$

For the purpose of guiding the inputs of self-attention modules, we downsample the optical flow V of all frames to the resolution of the hidden states.

4 Method

Given an input video sequence of n frames denoted as $L = [I_1, \dots, I_n]$ and a textual prompt τ describing the target edit, we aim to edit the video according to the textual prompt and generate a synthesized video sequence $J = [J_1, \dots, J_n]$ while preserving the global layout and local details. Directly applying a pre-trained text-to-image diffusion model ϵ_θ on each frame independently may lead to content inconsistencies across frames. These inconsistencies can be addressed by imposing temporal constraints to the features in the pre-trained diffusion model ϵ_θ .

Natural video is highly redundant in time and has a strong temporal consistencies across frames. Similar temporal consistencies are also present in the internal diffusion features of the input frames [11]. Based on this, we propose a flow-guided feature warp strategy to simultaneously constrain the query, key, and value of self-attention modules in ϵ_θ , thereby unifying the internal diffusion features.

Our framework, as shown in Fig. 3, has two stages: (i) DDIM inversion. (ii) Keyframe Joint editing and propagation. To preserve temporal consistencies across frames, we integrate our flow-guided feature warp strategy into both two stages. However, we find that directly using feature warp in DDIM inversion leads to error accumulation. As such, we propose a random shuffle strategy. In the following parts, we present the details of each stage and the corresponding strategy.

4.1 Flow-Guided Feature Warping Strategy

As mentioned in Sec. 3.3, the optical flow is computed by matching similar image contents and correspondences of the internal features between two consecutive RGB frames. Therefore, we propose to use optical flow of the source video to warp the input features of the residual blocks and the self-attention modules between two consecutive frames.

Let $I_n, I_{n+1} : P \rightarrow \mathbb{R}^3$ be two consecutive frames, as shown in the right part in Fig. 3, we first extract the forward optical flow $w^f = (u^f, v^f)^T$ and the

backward optical flow $w^b = (u^b, v^b)^T$. Since there exists occluded regions after the warping operation, we calculate an occlusion mask B denoting the occlusion regions of the warped features. Considering the binary mask generation, we refer to a previous work [24] and generate the occlusion mask based on the forward-backward consistency assumption [40].

For pixels that are not occluded, the forward flow should be the backward flow at the corresponding pixel in the next frame. Hence, a considerable disparity exists between the two flows of occluded pixels. We define the occlusion mask B_o to be 1 whenever the constraint

$$\begin{aligned} & |w^f(x) + w^b(x + w^f(x))|^2 \\ & < \alpha_1 (|w^f(x)|^2 + |w^b(x + w^f(x))|^2) + \alpha_2, \end{aligned} \quad (7)$$

is violated, and 0 otherwise. We set the hyper-parameters $\alpha_1 = 0.01$ and $\alpha_2 = 0.5$ for all $x \in P$.

The occlusion flag B_o addresses concerns with occlusion but falls short in handling scene transitions within video contexts. Therefore, we compute the residual map $r = |W(I_n, w^f) - I_{n+1}|$, where W denotes a backward warping operation. Taking into consideration of both the occlusion mask B_o and the residual map r into consideration, we define the final binary mask as:

$$B = \begin{cases} 1, & \text{if } B_o - \alpha r > \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

For frame $n + 1$, the mask separates the features that should be warped from frame n (value of 1) versus the features that should be generated (value of 0). With the backward flow w^b and the binary mask B , we perform a feature alignment process in the residual blocks and self-attention modules. In each denoising timestep, we save the input internal feature $\hat{z}_n^i$ of residual blocks or self-attention modules, where i denotes the timestep and n denotes the frame index. Afterwards, the internal feature $\hat{z}_{n+1}^i$ of frame $(n + 1)$ frame is computed by:

$$\hat{z}_{n+1}^i = B * w^b(\hat{z}_n^i) + (1 - B)z_{n+1}^i, \quad (8)$$

Here, z_{n+1}^i represents the feature generated from the $(n + 1)$ -th frame. The binary mask B is downsampled to the size of the internal feature. This process ensures the close generation of features within the residual blocks and the query, key, and value tokens of the self-attention modules, aligning them closely with related features from the previous frame in temporally corresponding regions. Consequently, maintaining a coherent update to the latent feature supports the continual preservation of visual consistency between frames throughout the sequence. This flow-guided feature warping strategy is integrated into the DDIM inversion and joint editing part.

4.2 Randomly Shuffled DDIM Inversion

Given an input video sequence L , DDIM inversion provides the inverted noise latent for video editing [39]. However, the inverted noise latents may result in

frame inconsistencies due to the error accumulation during the DDIM denoising steps and the impact of classifier-free guidance [41]. We suppose that severe errors occur at only a few timesteps per frame and accumulate throughout the entire inference process and can be offset using temporal information. To mitigate this error, we integrate our flow-guided feature warping strategy into the DDIM inversion process as shown in the left part of Fig. 3, where the purple box denotes our Flow-Guided Feature Warping Strategy.

Curiously, we noticed that a fixed order inversion leads to warping error accumulation. We speculate this is caused by errors introduced during the optical flow downscaling or the binary mask estimation. To address this issue, we first randomly invert the sequential order of the input frames. We then divide this sequence into small groups and randomly shuffle the frames in each group. The DDIM inversion process is then applied to the shuffled frames sequence based on the flow-guided feature warping strategy. The resulting inverted noise latents are denoted as $[x_n^i]$, where i is the denoise timestep and n is the frame index. Subsequently, these latent will be employed to establish inter-frame correspondences among diffusion features.

4.3 Keyframe Joint Editing

The joint editing process of the keyframes is performed on the inverted noise latents of the randomly shuffled DDIM inversion. We first integrate our flow-guided feature warp strategy into the editing process to constrain the query, key, and value in the self-attention modules. To further align the key and value of all the frames, we extend the self-attention block to the extended-attention block, which encourages all the edited frames to share a unified global appearance. Specifically, consider the input of self-attention features from all N keyframes $\{Q_n\}_{n \in \mathcal{K}}, \{K_n\}_{n \in \mathcal{K}}, \{V_n\}_{n \in \mathcal{K}}$, where Q_n , K_n and V_n are the queries, keys and values of the frame $n \in \mathcal{K}$, $\mathcal{K} = \{1, \dots, N\}$. We concatenate the keys and the values of each frame with the first frame and calculate the extended-attention as follows:

$$\mathbf{ExtAttn}(Q_n; [K_1, K_n]) = \mathbf{Softmax} \left(\frac{Q_n [K_1, K_n]^T}{\sqrt{d}} \right), \quad (9)$$

The output feature of n -th feature in each self-attention module is given by:

$$F_n = \hat{A} \cdot [V_1, V_n], \quad (10)$$

Intuitively, each keyframe queries the first frame and itself and aggregates the information. This leads to a roughly unified appearance in the edited frames.

4.4 Full Video Propagation

We sample the keyframes for every K frame, i.e., $I_0, I_K, I_{2K}, \dots$ and edit them using our unified framework. The shuffled edited keyframes are re-ordered in the correct sequence. To generate the full video, we employ Ebsynth [17], which

propagates the stylized keyframes to neighboring frames with a guided patch-matching algorithm that incorporates color, positional, edge, and temporal guidance. Specifically, for two keyframes I_0 and I_K and $(0 < i < K)$ in between, the two keyframes are warped to frame i to derive I_i^0 and I_i^K . Initially, Ebsynth combines the colors and gradients of I_i^0 and I_i^K by selecting the ones with minimal errors during patch matching for each location. Subsequently, the combined color image serves as a histogram reference for contrast-preserving blending [13] over I_i^0 and I_i^K to generate the final blended image of I_i .

5 Experiments

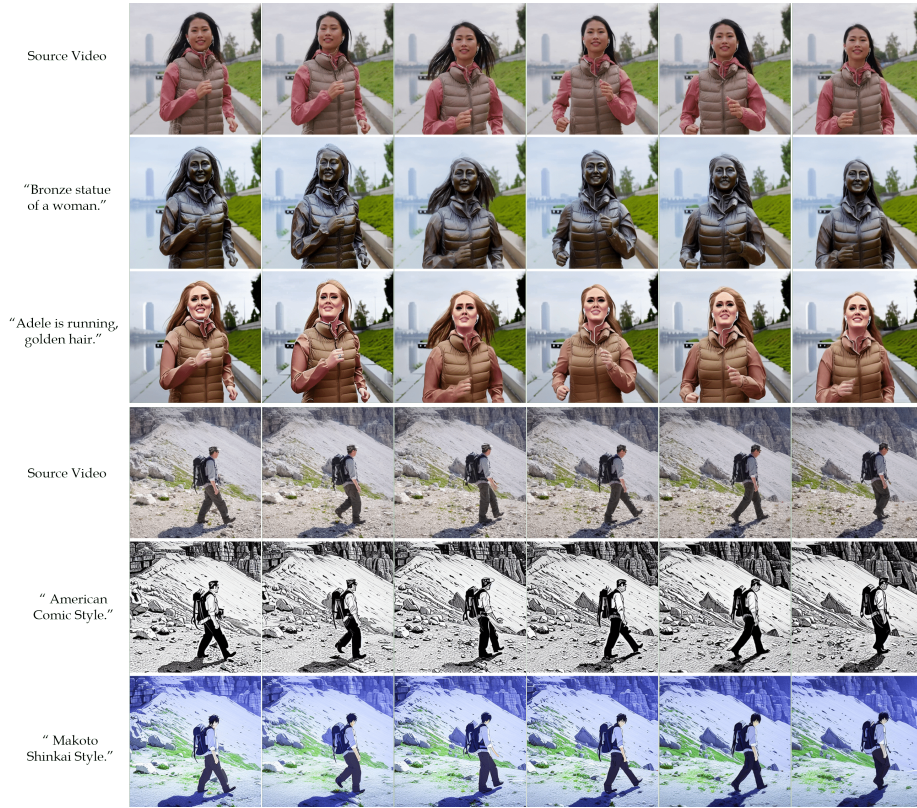


Fig. 4: Visualization results. The editing results of two videos with different prompts demonstrate that our method achieves high-quality editing and strong temporal coherence.

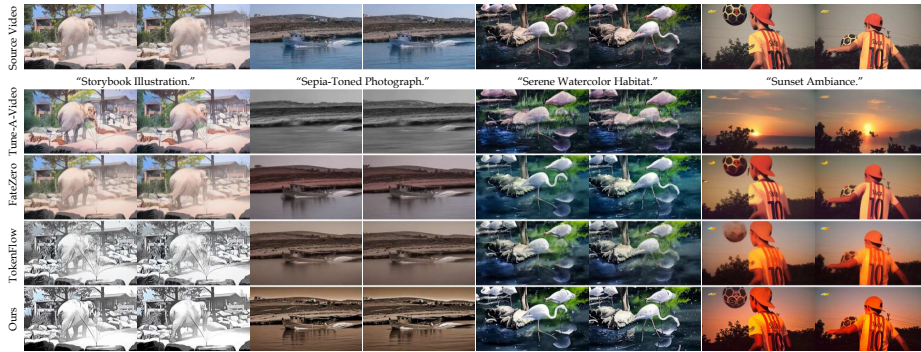


Fig. 5: Visualization. We reproduce these methods using their official codes and recommended parameters.

5.1 Implementation Details

Our implementation is based on Stable Diffusion [35] and their publicly pre-trained weights¹. We utilize PnP-Diffusion [41] as our frame editing method. Additionally, we incorporate ControlNet [52] with an edge condition² to more effectively harness the structural information from the source video. The number of DDIM inversion timesteps and denoising timesteps are set at $T = 50$. We adopt our flow-guided feature warping strategy in the first 40 timesteps and use the normal denoising strategy in the last 10. We set the interval between keyframes at 4 or 8. All the experiments are conducted with one NVIDIA A100 GPU.

Datasets. Following Tune-A-Video [44], we evaluate our method on 42 videos from DAVIS dataset [30]. We first use GPT-4 to generate text prompts and manually design 84 text-video pairs. Each video has a spatial resolution of 384×672 and is split into 24 frames. Our method is also tested on other Internet videos about animals, humans, and object-moving scenes for visualization. More details about our videos and prompts can be found in our supplementary materials (SM).

Baselines. We compare our method with three publicly available text-driven editing methods: Tune-A-Video [44], FateZero [31], and TokenFlow [11]. In these works, Tune-A-Video fine-tunes the inflated spatio-temporal attention layers on the source video. Furthermore, both FateZero and TokenFlow utilize a cross-frame attention module to preserve temporal consistency.

¹<https://huggingface.co/runwayml/stable-diffusion-v1-5>

²<https://huggingface.co/lllyasviel/sd-controlnet-canny>

Method	PSNR $\uparrow$	CLIP Score $\uparrow$	Warp-err $\downarrow$	LPIPS $\downarrow$
DDIM Inversion	23.02	22.04	64.06	0.2434
Ours	22.65	22.23	49.38	0.2330

Table 1: Quantitative evaluation. Comparison of the reconstruction quality of our randomly shuffled DDIM inversion and naive DDIM inversion using PSNR, CLIP, Warp error ($\times 10^{-3}$), and LPIPS.

5.2 Qualitative Comparison

Fig. 5 presents a qualitative comparison between our method and prominent baselines. Additional videos can be found in the SM. Tune-A-Video (second row) transforms a 2D image model into a video model and refines it to closely match the motion of the video, making it well-suited for short clips. However, when the textual prompt is relatively different from the input video, it fails to edit the video correctly (e.g., Sepia-Toned Photograph and Sunset Ambiance in Fig. 5). FateZero (third row) and TokenFlow (fourth row) produce results with commendable structural and inter-frame coherence, but its generated videos manifest blurring and temporal flicker (e.g., Storybook illustration and Serene Watercolor Habitat in Fig. 5). The videos generated by our method (bottom row) exhibit improved adherence to the edit prompt, while simultaneously preserving the temporal consistency of the edited video.

5.3 Quantitative Comparison

Our evaluation metrics include PSNR (peak signal-to-noise ratio), CLIP score, warp error, and LPIPS [53]. We also conduct user study, which can be found in the SM. The CLIP score assesses video-prompt alignment through average cosine similarity between CLIP text embeddings of the prompt and image embeddings of each frame. Warp error, measuring pixel-level consistency, computes the mean squared error of pixel values between warped edited frames and corresponding target frames. LPIPS evaluates the perceptual similarity between images. We first compare the reconstruction quality of our randomly shuffled inversion with naive DDIM inversion. Table 1 demonstrates our randomly shuffled inversion’s performance improvements in CLIP, Warp error, and LPIPS metrics by 0.19, 14.68, and 0.0104, respectively. Our PSNR result is slightly lower than that of naive DDIM inversion (22.65dB vs 23.02 dB), yet our method achieves more accurate reconstruction, faithfully preserving the source video’s content in each frame, as shown in Fig. 6. The comparison of editing results is provided in Table 2. Our proposed method achieves the least warp error (1.83 lower than TokenFlow), showing a good temporal consistency. Furthermore, we also achieve a higher CLIP score that indicates a good matching between the edited video and the input textual prompt. Our analysis reveals that Tune-A-Video achieves the highest CLIP score due to its generation of frames that are unfaithful to the original content but align closely with textual prompts, as illustrated in Fig. 5.

Method	CLIP Score $\uparrow$	Warp-err $\downarrow$
Tune-A-Video [44]	25.36	32.33
FateZero [31]	22.97	9.32
TokenFlow [11]	23.18	10.77
Ours w/ DDIM Inversion	24.33	9.71
Ours w/o Warping	23.60	9.89
Ours	24.44	8.94

Table 2: Quantitative evaluation. We evaluate our method in temporal consistency, and prompt alignments by computing warp-error($\times 10^{-3}$) and CLIP similarity, respectively.

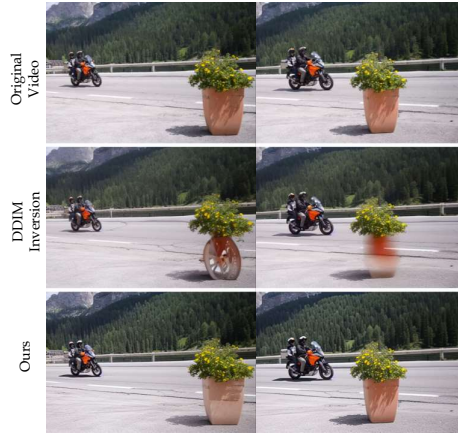


Fig. 6: Visualization. Comparison of reconstruction results between Naive DDIM inversion and ours.

5.4 Ablation Study

We conducted ablation studies on our randomly shuffled inversion and warping operation, focusing on their impact on maintaining temporal consistency. “Ours w/ DDIM Inversion” indicates the use of the naive DDIM inversion within our framework, while “Ours w/o Warping” signifies the removal of the warping operation only from our editing process. A quantitative comparison of the edited videos is detailed in Table 2, below the dotted line. We further visualize the edited results in Fig. 7. In the second row, the error, characterized by a blurry tail, accumulates into subsequent frames, resulting in structural misinformation generation (small grass). In the third row, the tail of the camel remains as our randomly shuffled DDIM inversion can keep the local details of the original image. However, the color of the background region has changed, which is denoted by the purple box. This shows that our flow-guided feature warp strategy can ensure temporal consistencies across frames.

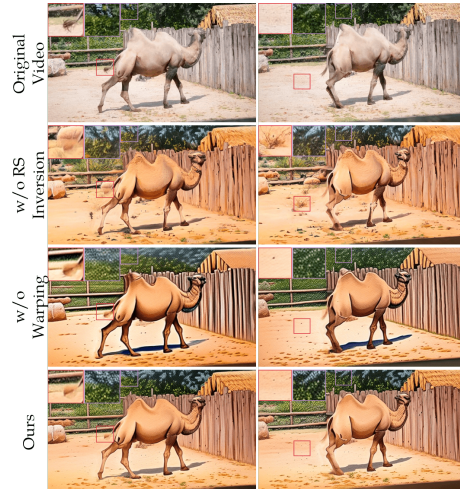


Fig. 7: Ablation study. We ablate the effect of our RS inversion and Warping operation through visualization.

6 Limitations and Conclusion

Limitations. As for limitations, while our approach endeavors to extract ample structural information from the original video to ensure temporal consistency and stability in the transferred video, there is still room for improvement. This method might exert overly stringent control on the video’s structure, thereby impeding the attainment of significant structural variations. Such limitations may hinder the creative potential of the editing process. Also, as noted in [11], the transferred video may encounter high-frequency flickering introduced by the LDM decoder. Meanwhile, our framework’s reliance on image editing technology imposes constraints on the efficacy of video editing. The extent of editing possibilities in the transferred video is bound by the capabilities of the corresponding image editing model, thus potentially limiting the range of effects and transformations that can be accomplished.

Conclusion. In this paper, we introduce a novel warping strategy within the feature domain for text-driven video editing based on current image diffusion model techniques. By integrating temporal keyframe information, we effectively reduce error accumulation inherent in the inversion process and achieve high-quality inverted latents for subsequent processes. Our unified zero-shot framework employs this warping strategy during both inversion and denoising processes. Experiments compared with existing state-of-the-art methods show that our edited videos can maintain temporal coherence in both global style and local textures.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its NRF Fellowship for AI (NRF-NRFFAI1-2019-0001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. The work is supported by the National Key Research and Development Program of China (Grant No. 2023YFC3300029). This research is also supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016 and 12201024, "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268, Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park Grant No.Z231100005923035. Taiyuan City "Double hundred Research action" 2024TYJB0127.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. arXiv preprint arXiv:2206.02779 (2022)
2. Bao, Y., Qiu, D., Kang, G., Zhang, B., Jin, B., Wang, K., Yan, P.: Latentwarp: Consistent diffusion latents for zero-shot video-to-video translation. arXiv preprint arXiv:2311.00353 (2023)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR. pp. 22563–22575 (2023)
4. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. arXiv preprint arXiv:2303.12688 (2023)
5. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: Glean: Generative latent bank for large-factor image super-resolution. In: CVPR. pp. 14245–14254 (2021)
6. Cong, Y., Xu, M., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S., et al.: Flatten: optical flow-guided attention for consistent text-to-video editing. In: ICLR (2023)
7. Dong, W., Xue, S., Duan, X., Han, S.: Prompt tuning inversion for text-driven image editing using diffusion models. In: ICCV. pp. 7430–7440 (2023)
8. Duan, X., Cui, S., Kang, G., Zhang, B., Fei, Z., Fan, M., Huang, J.: Tuning-free inversion-enhanced control for consistent image editing. In: AAAI. vol. 38, pp. 1644–1652 (2024)
9. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. arXiv preprint arXiv:2302.03011 (2023)
10. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: CVPR. pp. 10021–10030 (2023)
11. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: ICLR (2024)
12. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. In: CVPR. pp. 3012–3021 (2020)

13. Heitz, E., Neyret, F.: High-performance by-example noise using a histogram-preserving blending operator. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* **1**(2), 1–25 (2018)
14. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *NeurIPS* **33**, 6840–6851 (2020)
16. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *arXiv:2204.03458* (2022)
17. Jamriška, O., Sochorová, Š., Texler, O., Lukáč, M., Fišer, J., Lu, J., Shechtman, E., Šykora, D.: Stylizing video by example. *ACM TOG* **38**(4), 1–11 (2019)
18. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: *ICCV*. pp. 15954–15964 (2023)
19. Li, H., Feng, Y., Xue, S., Liu, X., Zeng, B., Li, S., Liu, B., Liu, J., Han, S., Zhang, B.: Uv-idm: Identity-conditioned latent diffusion model for face uv-texture generation. In: *CVPR*. pp. 10585–10595 (2024)
20. Li, S., Zeng, B., Feng, Y., Gao, S., Liu, X., Liu, J., Li, L., Tang, X., Hu, Y., Liu, J., et al.: Zone: Zero-shot instruction-guided local editing. In: *CVPR*. pp. 6254–6263 (2024)
21. Liang, J., Pei, W., Lu, F.: Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In: *ECCV*. pp. 491–508. Springer (2020)
22. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761* (2023)
23. Liu, X., Zeng, B., Gao, S., Li, S., Feng, Y., Li, H., Liu, B., Liu, J., Zhang, B.: LadiFgan: Training gans with diffusion supervision in latent spaces. In: *CVPRW*. pp. 1115–1125 (2024)
24. Meister, S., Hur, J., Roth, S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In: *AAAI*. vol. 32 (2018)
25. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: *ICLR* (2022)
26. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794* (2022)
27. Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., Hoshen, Y.: Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329* (2023)
28. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: *ICML*. pp. 16784–16804. PMLR (2022)
29. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. In: *CVPR*. pp. 14104–14113 (2020)
30. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017)
31. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: *ICCV*. pp. 15932–15942 (2023)
32. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: *CVPR*. pp. 1505–1514 (2019)

33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022)
34. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML. pp. 1060–1069. PMLR (2016)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
36. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
37. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* **35**, 36479–36494 (2022)
38. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: ICLR (2023)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
40. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by gpu-accelerated large displacement optical flow. In: ECCV. pp. 438–451. Springer (2010)
41. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR. pp. 1921–1930 (2023)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* **30** (2017)
43. Viazovetskyi, Y., Ivashkin, V., Kashin, E.: Stylegan2 distillation for feed-forward image manipulation. In: ECCV. pp. 170–186. Springer (2020)
44. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV. pp. 7623–7633 (2023)
45. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. *IEEE TPAMI* **45**(3), 3121–3138 (2022)
46. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: CVPR. pp. 8121–8130 (2022)
47. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: CVPR. pp. 18456–18466 (2023)
48. Zeng, B., Li, S., Feng, Y., Li, H., Gao, S., Liu, J., Li, H., Tang, X., Liu, J., Zhang, B.: Ipdreamer: Appearance-controllable 3d object generation with image prompts. *arXiv preprint arXiv:2310.05375* (2023)
49. Zeng, B., Li, S., Liu, X., Gao, S., Jiang, X., Tang, X., Hu, Y., Liu, J., Zhang, B.: Controllable mind visual diffusion model. In: AAAI. vol. 38, pp. 6935–6943 (2024)
50. Zeng, B., Liu, X., Gao, S., Liu, B., Li, H., Liu, J., Zhang, B.: Face animation with an attribute-guided diffusion model. In: CVPRW. pp. 628–637 (2023)
51. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stack-GAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV. pp. 5907–5915 (2017)
52. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
54. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098* (2023)

55. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
56. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV. pp. 597–613. Springer (2016)
57. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR. pp. 5802–5810 (2019)