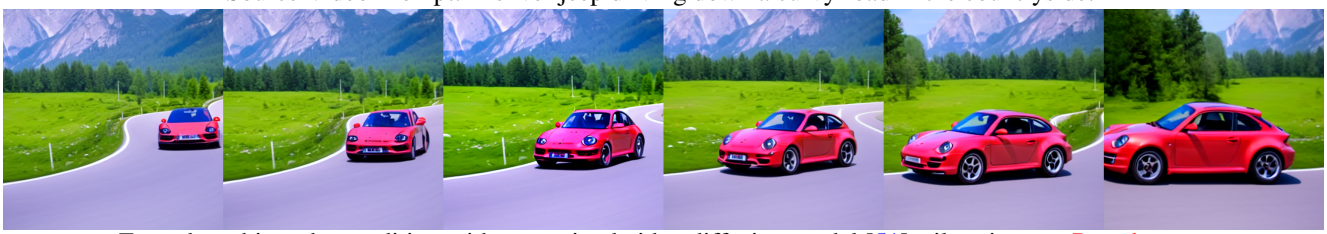
FateZero: Fusing Attentions for Zero-shot Text-based Video Editing

Chenyang Qi[1*] Xiaodong Cun[2†] Yong Zhang[2] Chenyang Lei[3]
Xintao Wang[2] Ying Shan[2] Qifeng Chen[1†]

[1]HKUST [2]Tencent AI Lab [3]CAIR, HKISI-CAS

https://fate-zero-edit.github.io

Source Video Prompt: A silver jeep driving down a curvy road in the countryside.



Zero-shot object shape editing with pre-trained video diffusion model [51]: silver jeep → Porsche car.

Zero-shot video style editing with pre-trained image diffusion model [41]: watercolor painting.

Figure 1. **Zero-shot text-driven video editing.** We present a zero-shot approach for shape-aware local object editing and video style editing from pre-trained diffusion models [41, 51] without any optimization for each target prompt.

Abstract

The diffusion-based generative models have achieved remarkable success in text-based image generation. However, since it contains enormous randomness in generation progress, it is still challenging to apply such models for real-world visual content editing, especially in videos. In this paper, we propose `FateZero`*, a zero-shot text-based editing method on real-world videos without per-prompt training or use-specific mask. To edit videos consistently, we propose several techniques based on the pre-trained models. Firstly, in contrast to the straightforward DDIM inversion technique, our approach captures intermediate attention maps during inversion, which effectively retain both structural and motion information. These maps are directly fused in the editing process rather than generated during denoising. To further minimize semantic leakage of the source video, we then fuse self-attentions with a blending*

* Work done during an internship at Tencent AI Lab.
† Corresponding Authors.

mask obtained by cross-attention features from the source prompt. Furthermore, we have implemented a reform of the self-attention mechanism in denoising UNet by introducing spatial-temporal attention to ensure frame consistency. Yet succinct, our method is the first one to show the ability of zero-shot text-driven video style and local attribute editing from the trained text-to-image model. We also have a better zero-shot shape-aware editing ability based on the text-to-video model [51]. Extensive experiments demonstrate our superior temporal consistency and editing capability than previous works.

1. Introduction

Diffusion-based models [19] can generate diverse and high-quality images [39, 41, 43] and videos [15, 18, 44, 55] through text prompts. It also brings large opportunities to edit real-world visual content from these generative priors.

Previous or concurrent diffusion-based editing methods [2, 3, 6, 16, 37, 47] majorly work on images. To edit real images, their methods utilize deterministic DDIM [45] for the image-to-noise inversion, and then, the inverted noise gradually generates the edited images under the condition of the target prompt. Based on this pipeline, several methods have been proposed in terms of cross-attention guidance [37], plug-and-play feature [47], and optimization [25, 34].

Manipulating videos through generative priors as image editing methods above contains many challenges (Fig. 7). First, there are no publicly available generic text-to-video models [18, 44]. Thus, a framework based on image models can be more valuable than on video ones [35], thanks to the various open-sourced image models in the community [1, 36, 41, 53]. However, the text-to-image models [41] lack the consideration of temporal-aware information, *e.g.*, motion and 3D shape understanding. Directly applying the image editing methods [32, 34] to the video will show obverse flickering. Second, although we can use previous video editing methods [4, 24, 28] via keyframe [21] or atlas editing [4, 24], these methods still need atlas learning [4, 24], keyframe selection [21], and per-prompt tunning [4, 28]. Moreover, while they may work well on the attribute [4, 24] and style [21] editing, the shape editing is still a big challenge [28]. Finally, as introduced above, current editing methods use DDIM for inversion and then denoising via the new prompt. However, in video inversion, the inverted noise in the T step might break the motion and structure of the original video because of error accumulation (Fig. 4 and 9).

In this paper, we propose `FateZero`, a simple yet effective method for zero-shot video editing since we do not need to train for each target prompt individually [4, 24, 28] and have no user-specific mask [2, 3]. Different from image editing, video editing needs to keep the temporal consistency of the edited video, which is not learned by the original trained text-to-image model. We tackle this problem by using two novel designs. Firstly, instead of solely relying on inversion and generation [16, 34, 47], we adopt a different approach by storing all the self and cross-attention maps at every step of the inversion process. This enables us to subsequently replace them during the denoising steps of the DDIM pipeline. Specifically, we find these self-attention blocks store better motion information and the cross-attention can be used as a threshold mask for self-attention blending spatially. This attention blending operation can keep the original structures unchanged. Furthermore, we reform the self-attention blocks to the spatial-temporal attention blocks as in [51] to make the appearance more consistent. Powered by our novel designs, we can directly edit the style and the attribute of the real-world video (Fig. 6) using the pre-trained text-to-image model [41]. Also, after getting the video diffusion model (*e.g.*, pretrained Tune-A-Video [51]), our method shows better object editing (Fig. 5) ability in test-time than simple DDIM inversion [45]. The extensive experiments provide evidence of the advantages offered by the proposed method for both video and image editing.

Our contributions are summarized as follows:

- We present the first framework for temporal-consistent zero-shot text-based video editing using pretrained text-to-image model.
- We propose to fuse the attention maps in the inversion process and generation process to preserve the motion and structure consistency during editing.
- Our novel Attention Blending Block utilizes the source prompt's cross-attention map during attention fusion to prevent source semantic leakage and improve the shape-editing capability.
- We show extensive applications of our method in video style editing, video local editing, video object replacement, *etc*.

2. Related Work

Video Editing. Video can be edited via several aspects. For video stylizing editing, current methods [11, 21] rely on the example as the style guide and these methods may fail when the track is lost. By processing frames individually using image style transfer [13, 22], some works also learn to reduce the temporal consistency [5, 27, 29, 30] in a post-process way. However, the style may still be imperfect since the style transfer only measures the perceptual distance [54]. Several works also show better consistency but on the specific domain, *e.g.*, portrait video [12, 52]. For video local editing, layer-atlas based methods [4, 24] show a promising direction by editing the video on a flattened texture map. However, the 2d atlas lacks 3d motion perception to support shape editing, and prompt-specific optimization is required.

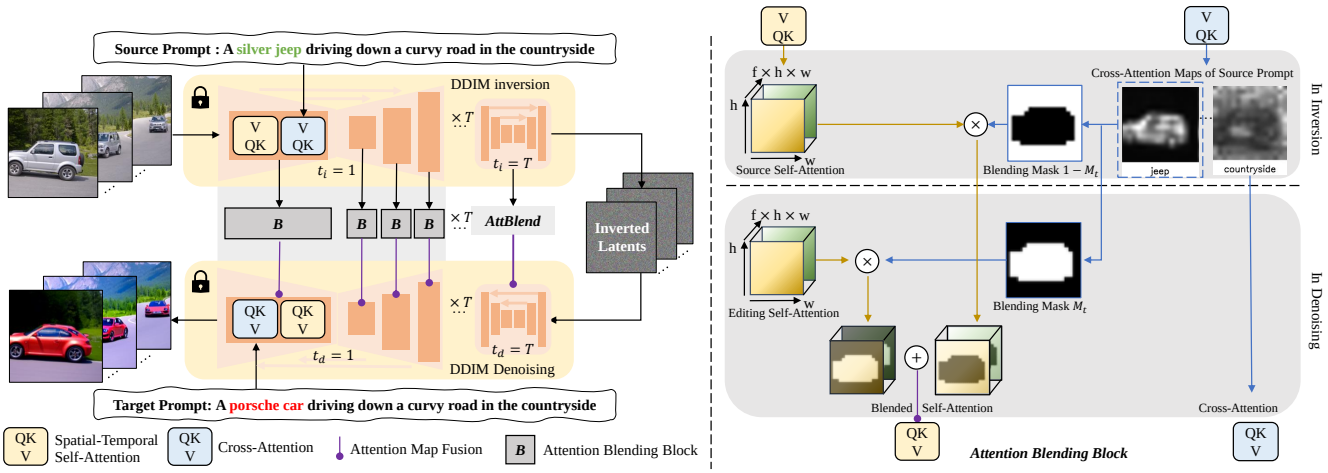


Figure 2. **The overview of our approach**. Our input is the user-provided source prompt p_{src}, target prompt p_{edit} and clean latent $z = \{z^1, z^2, ...z^n\}$ encoded from input source video $x = \{x^1, x^2, ...x^n\}$ with number frames n in a video sequence. On the **left**, we first invert the video using DDIM inversion pipeline into noisy latent z_T using the source prompt p_{src} and an inflated 3D U-Net ε_θ. During each inversion timestep t, we store both spatial-temporal self-attention maps s_t^{src} and cross-attention maps c_t^{src}. At the editing stage of the DDIM denoising, we denoise the latent z_T back to clean image $\hat{z}_0$ conditioned on target prompt p_{edit}. At each denoising timestep t , we fuse the attention maps (s_t^{edit} and c_t^{edit}) in ε_θ with stored attention map (s_t^{src}, c_t^{src}) using the proposed Attention Blending Block. **Right**: Specifically, we replace the cross-attention maps c_t^{edit} of un-edited words (*e.g.*, road and countryside) with source maps c_t^{src} of them. In addition, we blend the self-attention map during inversion s_t^{src} and editing s_t^{edit} with an adaptive spatial mask obtained from cross-attention c_t^{src}, which represents the areas that the user wants to edit.

A more challenging topic is to edit the object shape in the real-world video. Current method shows obvious artifacts even with the optimization on generative priors [28]. The stronger prior of the diffusion-based model also draws the attention of current researchers. *e.g.*, gen1 [9] trains a conditional model for depth and text-guided video generation, which can edit the appearance of the generated images on the fly. Dreamix [35] finetunes a stronger diffusion-based video model [18] for editing with stronger generative priors. Both of these methods need privacy and powerful video diffusion models for editing. Thus, the applications of the current larger-scale fine-tuned text-to-image models [1] cannot be used directly.

Image and Video Generation Models. Image generation is a basic and hot topic in computer vision. Early works mainly use VAE [26] or GAN [14] to model the distribution on the specific domain. Recent works adopt VQVAE [48] and transformer [10] for image generation. However, due to the difficulties in training these models, they only work well on the specific domain, *e.g.*, face [23]. On the other hand, the editing ability of these models is relatively weak since the feature space of GAN is high-level, and the quantified tokens can not be considered individually. Another type of method focuses on text-to-image generation. DALL-E [39, 40] and CogView [8] train an image generative pre-training transformer (GPT) to generate images from a CLIP [33] text embedding. Recent models [41, 43] benefit from the stability of training diffusion-based model [19]. These models can be scaled by a huge dataset and show surprisingly good results on text-to-image generation by integrating large language model conditions since its latent space has spatial structure, which provides a stronger edit ability than previous GAN [23] based methods. Generating videos is much more difficult than images. Current methods rely on the larger cascaded models [18, 44] and dataset. Differently, magic-video [55] and gen1 [9] initialize the model from text-to-image [41] and generate the continuous contents through extra time-aware layers. Recently, Tune-A-Video [51] overfits a single video for text-based video generation. After training, the model can generate related motion from similar prompts. However, how to edit real-world content using this model is still unclear. Inspired by the image editing methods and tune-a-video, our method can edit the style of the real-world video and images using the trained text-to-image model [41] and shows better object replacing performance than the one-shot finetuned video diffusion model [51] with simple DDIM inversion [45] in real videos (Fig. 7).

Image Editing in Diffusion Model. Many recent works adopt the trained diffusion model for editing. SDEdit [32] generates content for a new prompt by adding noise to the image first. DiffEdit [6] computes the edit mask by the noise differences of the text prompts, and then, blends the inversion noises into the image generation process. Similar work has also been proposed by Blended Diffusion [2, 3], which combines the features of each step for image blending. Plug-and-play [47] gets the inversion noise and applies the denois-

ing for feature reconstruction. After that, the self-attention features in editing are replaced with that in reconstruction directly. Pix2pix-Zero [37] edits the image with the cross-attention guidance. Prompt-to-Prompt [16] proves that images can be edited via reweighting the cross-attention map of different prompts. There are also some methods to achieve better editing ability via optimization [25, 34]. However, a naive frame-wise application of these image methods to video results in flickering and inconsistency among frames.

3. Methods

We target zero-shot text-driven video editing (*e.g.*, style, attribute, and shape) without optimization for each target prompt or the user-provided mask. In Sec. 3.1, we first give the details of the latent diffusion and DDIM inversion. After that, we introduce our method that enables video appearance editing (Sec. 3.2) via the pre-trained text-to-image models [41]. Finally, we discuss a more challenging case that also enables the shape-aware editing of video using the video diffusion model in Sec. 3.3. Notice that, the proposed method is a general editing method and can be used in various text-to-image or text-to-video models. In this paper, we majorly use Stable Diffusion [41] and the video generation model based on Stable Diffusion (Tune-A-Video [51]) for its popularity and generalization ability.

3.1. Preliminary: Latent Diffusion and Inversion

Latent Diffusion Models [41] are introduced to diffuse and denoise the latent space of an autoencoder. First, an encoder $\mathcal{E}$ compresses a RGB image x to a low-resolution latent $z = \mathcal{E}(x)$, which can be reconstructed back to image $\mathcal{D}(z) \approx x$ by decoder $\mathcal{D}$. Second, a U-Net [42] ε_θ containing cross-attention and self-attention [49] is trained to remove the artificial noise using the objective:

$$\min_{\theta} E_{z_0,\varepsilon\sim N(0,I),t\sim \text{ Uniform }(1,T)} \left\|\varepsilon - \varepsilon_\theta\left(z_t, t, p\right)\right\|_2^2, \quad (1)$$

where p is the embedding of the conditional text prompt and z_t is a noisy sample of z_0 at timestep t.

DDIM Inversion [45]. During inference, deterministic DDIM sampling is employed to convert a random noise z_T to a clean latent z_0 in a sequence of timestep $t : T \to 1$:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \frac{z_t - \sqrt{1-\alpha_t}\varepsilon_\theta}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\varepsilon_\theta, \quad (2)$$

where α_t is a parameter for noise scheduling [19, 45]

Based on the ODE limit analysis of the diffusion process, DDIM inversion [7, 45] is proposed to map a clean latent z_0 back to a noised latent $\hat{z}_T$ in revered steps $t : 1 \to T$:

$$\hat{z}_t = \sqrt{\alpha_t} \frac{\hat{z}_{t-1} - \sqrt{1-\alpha_{t-1}}\varepsilon_\theta}{\sqrt{\alpha_{t-1}}} + \sqrt{1-\alpha_t}\varepsilon_\theta. \quad (3)$$

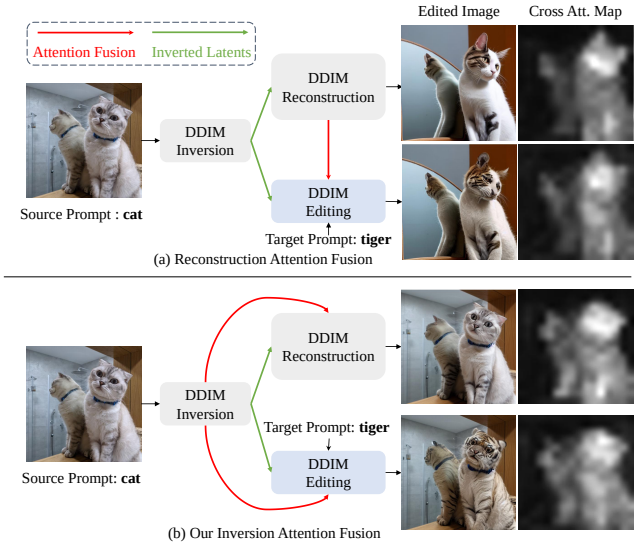


Figure 3. **Zero-shot local attributed editing (cat → tiger) using stable diffusion.** In contrast to fusion with attention during reconstruction (a) in previous work [16, 37, 47], our inversion attention fusion (b) provides more accurate structure guidance and editing ability, as visualized on the right side.

Such that the inverted latent $\hat{z}_T$ can reconstruct a latent $\hat{z}_0(p_{src}) = \text{DDIM}(\hat{z}_T, p_{src})$ similar to the clean latent z_0 at classifier-free guidance scale $s_{cfg} = 1$. Recently, image editing methods [16, 34, 37, 47] use a large classifier-free guidance scale $s_{cfg} \gg 1$ to edit the latent as $\hat{z}_0(p_{edit}) = \text{DDIM}(\hat{z}_T, p_{edit})$ (second row in Fig 3(a)), where a reconstruction of $\hat{z}_0(p_{src})$ is conducted in parallel to provide attention constraints. (first row in Fig 3(a)).

3.2. FateZero Video Editing

As shown in Fig. 2, we use the pretrained text-to-image model, *i.e.*, Stable Diffusion, as our base model, which contains a UNet for T-timestep denoising. Instead of straightforwardly exploiting the regular pipeline of latent editing guided by reconstruction attention, we have made several critical modifications for video editing as follows.

Inversion Attention Fusion. Direct editing using the inverted noise results in frame inconsistency, which may be attributed to two factors. First, the invertible property of DDIM discussed in Eq. (2) and Eq. (3) only holds in the limit of small steps [45, 46]. Nevertheless, the present requirements of 50 DDIM denoising steps lead to an accumulation of errors with each subsequent step. Second, using a large classifier-free guidance $s_{cfg} \gg 1$ can increase the edit ability in denoising, but the large editing freedom leads to inconsistent neighboring frames. Therefore, previous methods require optimization of text-embedding [16] or other regularization [37].

While the issues seem trivial in the context of single-

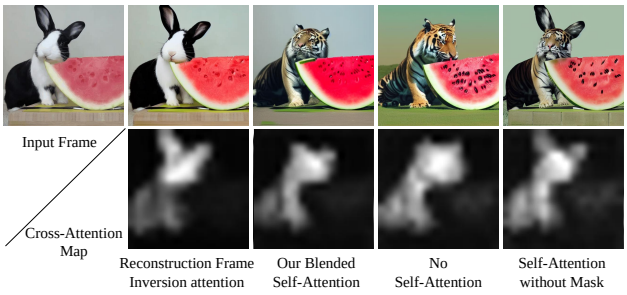


Figure 4. **Study of blended self-attention in zero-shot shape editing (rabbit → tiger) using stable diffusion**. Forth and fifth columns: Ignoring self-attention can not preserve the original structure and background, and naive replacement causes artifacts. Third column: Blending the self-attention using the cross-attention map (the second row) obtains both new shape from the target text with a similar pose and background from the input frame.

frame editing they can become magnified when working with video as even minor discrepancies among frames will be accentuated along the temporal indexes.

To alleviate these issues, our framework utilizes the attention maps during inversion steps (Eq. (3)), which is available because the source prompt p_{src} and initial latent z_0 are provided to the UNet during inversion. Formally, during inversion, we store the intermediate self-attention maps $[s_t^{\text{src}}]_{t=1}^T$, cross-attention maps $[c_t^{\text{src}}]_{t=1}^T$ at each timestep t and the final latent feature maps z_T as

$$z_T, [c_t^{\text{src}}]_{t=1}^T, [s_t^{\text{src}}]_{t=1}^T = \text{DDIM-INV}(z_0, p_{src}), \quad (4)$$

where DDIM-INV stands for the DDIM inversion pipeline discussed in Eq. (3). During the editing stage, we can obtain the noise to remove by fusing the attention from inversion:

$$\hat{\epsilon}_t = \text{ATT-FUSION}(\varepsilon_\theta, z_t, t, p_{\text{edit}}, c_t^{\text{src}}, s_t^{\text{src}}). \quad (5)$$

where p_{edit} represents the modified prompt. In function ATT-FUSION, we inject the cross-attention maps of the unchanged part of the prompt similar to Prompt-to-Prompt [16]. We also replace self-attention maps to preserve the original structure and motion during the style and attribute editing.

Fig. 3 shows a toy comparison example between our attention fusion method and the typical method with simply inversion and then generation as in [16, 34] for image editing. The cross-attention map during inversion captures the silhouette and the pose of the cat in the source image, but the map during reconstruction has a noticeable difference. While in the video, the attention consistency might influence the temporal consistency as shown in Fig. 8. This is because the spatial-temporal self-attention maps represent the correspondence between frames and the temporal modeling ability of existing video diffusion model [51] is not satisfactory.

Attention Map Blending. Inversion-time attention fusion might be insufficient in local attrition editing, as shown in an image example in Fig. 4. In the third column, replacing self-attention $s^{edit} \in \mathbb{R}^{hw \times hw}$ with s^{src} brings unnecessary structure leakage and the generated image has unpleasant blending artifacts in the visualization. On the other hand, if we keep s^{edit} during the DDIM denoising pipeline, the structure of the background and watermelon has unwanted changes, and the pose of the original rabbit is also lost. Inspired by the fact that the cross-attention map provides the semantic layout of the image [16], as visualized in the second row of Fig. 4, we obtain a binary mask M_t by thresholding the cross-attention map of the edited words during inversion by a constant τ [2, 3]. Then, the self-attention maps of editing stage s_t^{edit} and inversion stage s_t^{src} are blended with the binary mask M_t, as illustrated in Fig. 2. Formally, the attention map fusion is implemented as

$$M_t = \text{HEAVISIDESTEP}(c_t^{src}, \tau), \quad (6)$$

$$s_t^{\text{fused}} = M_t \odot s_t^{\text{edit}} + (1 - M_t) \odot s_t^{\text{src}}. \quad (7)$$

Spatial-Temporal Self-Attention. The previous two designs make our method a strong editing method that can preserve the better structure, and also a big potential in video editing. However, denoising each frame individually still produces inconsistent video. Inspired by the casual self-attention [15, 20, 49, 50] and recent one-shot video generation method [51], we reshape the original self-attention to Spatial-Temporal Self-Attention without changing pretrained weights. Specifically, we implement $\text{ATTENTION}(Q, K, V)$ for feature z^i at temporal index $i \in [1, n]$ as

$$Q = W^Q \mathbf{z}^i, K = W^K \left[\mathbf{z}^i; \mathbf{z}^{\text{w}}\right], V = W^V \left[\mathbf{z}^i; \mathbf{z}^{\text{w}}\right], \quad (8)$$

where $[\cdot]$ denotes the concatenation operation and W^Q, W^K, W^V are the projection matrices from pretrained model. Empirically, we find it is enough to warp the middle frame $\mathbf{z}^{\text{w}} = z^{\text{Round}[\frac{n}{2}]}$ for attribute and style editing. Thus, the spatial-temporal self-attention map is represented as $s_t^{src} \in R^{hw \times fhw}$, where $f = 2$ is the number of frames used as key and value. It captures both the structure of a single frame and the temporal correspondence with the warped frames.

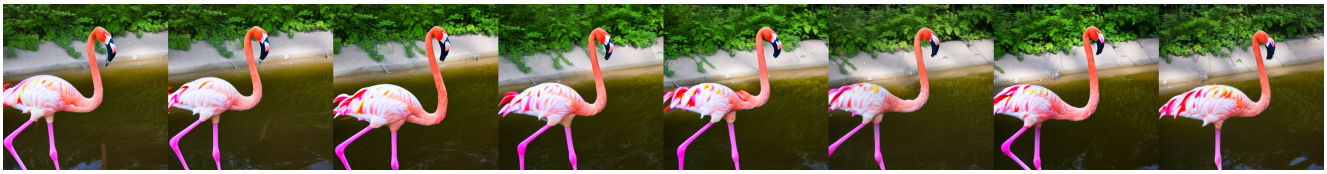
Overall, the proposed method produces a new editing method for zero-shot real-world video editing. We replace the attention maps in the denoising steps with their corresponding maps during the inversion steps. After that, we utilize cross-attention maps as masks to prevent semantic leaks. Finally, we reform the self-attention of UNet to spatial-temporal attention for better temporal consistency among different temporal frames. We have included a formal algorithm in the supplementary materials for reference purposes.

3.3. Shape-Aware Video Editing

Different from appearance editing, reforming the shape of a specific object in the video is much more challenging.

Source Prompt: A black swan with a red beak swimming in a river near a wall and bushes.

black swan → white duck.

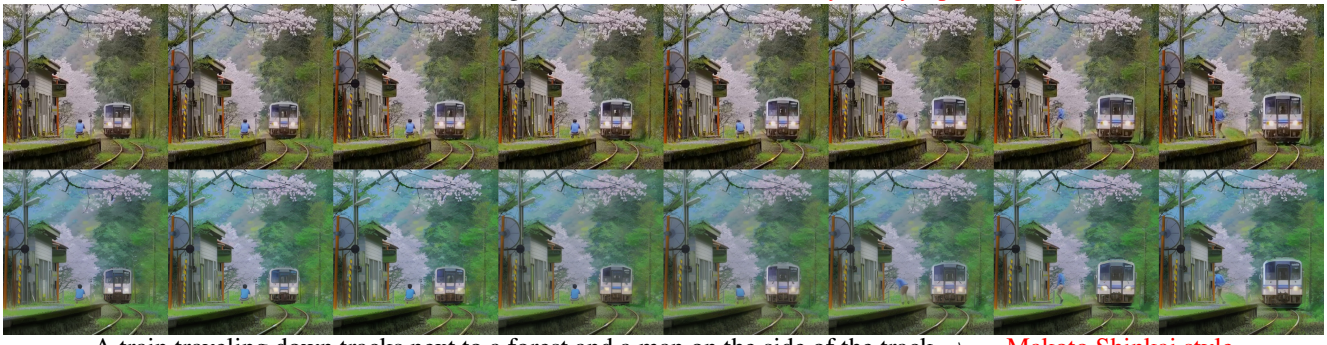


black swan → pink flamingo.

Figure 5. **Zero-shot object shape editing on pre-trained video diffusion model [51]:** Our framework can directly edit the shape of the object in videos driven by text prompts using a trained video diffusion model [51]

Source Prompt from Fig 5: black → Swarovski crystal

A man with round helmet surfing on a white wave → The Ukiyo-e style painting of a man ...



A train traveling down tracks next to a forest and a man on the side of the track → ..., Makoto Shinkai style

Figure 6. **Zero-shot attribute and style editing results using Stable Diffusion [41].** Our framework supports abstract attribute and style editing like 'Swarovski crystal', 'Ukiyo-e', and 'Makoto Shinkai'. Best viewed with zoom-in.

Figure 7. **Qualitative comparison of our methods with other baselines.** Inputs are in Fig. 5 and Fig 6. Our results have the best temporal consistency, image fidelity, and editing quality. Best viewed with zoom-in.

To this end, a pretrained video diffusion model is needed. Since there is no publicly-available generic video diffusion model, we perform the editing on the one-shot video diffusion model [51] instead. In this case, we compare our editing method with simple DDIM inversion [45], where our method also achieves better performance in terms of editing ability, motion consistency, and temporal consistency. It might be because it is hard for an inflated model to overfit the exact motion of the input video. While in our method, the motion and structure are represented by high-quality spatial-temporal attention maps $s_t^{src} \in R^{hw \times fhw}$ during inversion, which is further fused with the attention maps during editing. More details can be founded in Fig. 7 and the supp. video.

Method	CLIP Metrics↑		User Study↓		
Inversion & Editing	Tem-Con	Fram-Acc	Edit	Image	Temp
Framewise Null & p2p [16, 34]	0.852	**0.958**	3.55	4.11	4.38
Framewise SDEit [32]	0.910	0.819	3.69	3.28	3.62
NLA, Null & p2p [16, 24, 34]	0.949	0.600	3.17	3.02	2.60
Tune-A-Video & DDIM [45, 51]	0.958	0.750	2.78	2.80	2.70
Ours	**0.965**	0.903	**1.82**	**1.79**	**1.69**

Table 1. **Quantitative evaluation against baselines.** In our user study, the results of our method are preferred over those from baselines. For CLIP-Score, we achieve the best temporal consistency and comparable framewise editing accuracy against an optimization-based image editing method [34].

4. Experiments

4.1. Implementation Details

For zero-shot style and attribute editing, we directly use the trained stable diffusion v1.4 [41] as the base model, we fuse the attentions in the interval of $t \in [0.2 \times T, T]$ of the DDIM step with total timestep $T = 50$. For shape editing, we utilize the pretrained model of the specific video [51] at 100 iterations and fuse the attention at DDIM timestep $t \in [0.5 \times T, T]$, giving more freedom for new shape generation. Following previous works [4, 9], we use videos from DAVIS [38] and other in-the-wild videos to evaluate our approach. The source prompt of the video is generated via the image caption model [31]. Finally, we design the target prompt for each video by replacing or adding several words.

4.2. Applications

Local attribute and global style editing. Using pretrained text-to-image diffusion model [41], our framework supports zero-shot local attribute and global style editing, as shown in Fig. 6 and third row in Fig.1. In the first row, the texture and color of the feather are modified by the target prompt `Swarovski crystal` and kept consistent across frames. In the second and third rows, our framework applies abstract style (`Ukiyo-e` and `Makoto Shinkai`). The image structure and temporal motion can be well preserved since we fuse both the spatial-temporal self-attention and cross-attention during the inversion and editing stage.

Shape-aware editing. Fig. 5 and the second row in Fig.1 present the result of difficult object shape editing, with a pretrained video model [51]. This task is challenging because a naive full-resolution fusion of the spatial-temporal self-attention maps results in inaccurate shape results and wrong temporal motion, as shown in the ablation (Fig.9). Thanks to the proposed Attention Blending, we combine the motion of generated shape from the editing target and inverted attention from the input video. Results of `posche`, `duck` and `flamingo` show that we generate new content with poses and positions similar to input videos.

Zero-shot image editing. In addition, our framework can serve as a zero-shot image editing method such as local attribute editing (Fig. 3) and object shape editing (Fig. 4) by considering an image as a video with a single frame. We provide more results in our supplementary material.

4.3. Baseline Comparisons

Since there are no available zero-shot video editing methods based on diffusion models, we build the following four state-of-the-art baselines for comparison. (1) Tune-A-

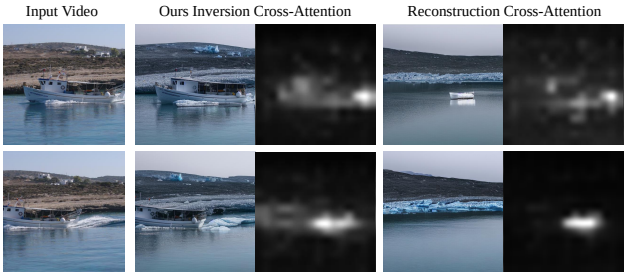


Figure 8. **Inversion attention compared with reconstruction attention using prompt 'deserted shore → 'glacier shore'.** The attention maps obtained from the reconstruction stage fail to detect the boat's position, and can not provide suitable motion guidance for zero-shot video editing.

Figure 9. **Ablation study of blended self-attention.** Without self-attention fusion, the generated video can not preserve the details of input videos (e.g., fence, trees, and car identity). If we replace full self-attention without a spatial mask, the structure of the original jeep misleads the generation of the Porsche car.

Video [51] overfits an inflated diffusion model on a single video to generate similar content. (2) The Neural Layered Atlas [24] (NLA) based method is combined with keyframe-editing via state-of-the-art image editing methods [16, 34]. (3) Frame-wise Null-text optimization [34] and then edit by prompt2prompt [16]. (4) Frame-wise zero-shot editing using SDEdit [32]. For attention-based editing (2,3,4), we use the same timesteps fusion parameters as ours.

We conduct the quantitative evaluation using the trained CLIP [33] model as previous methods [9, 37, 51]. Specially, we show the **'Tem-Con'** [9] to measure the temporal consistency in frames by computing the cosine similarity between all pairs of consecutive frames. **'Frame-Acc'** [17, 33, 37] is the frame-wise editing accuracy, which is the percentage of frames where the edited image has a higher CLIP similarity to the target prompt than the source prompt. In addition, three user studies metrics (denoted as **'Edit'**, **'Image'**, and **'Temp'**) are conducted to measure the editing quality, overall frame-wise image fidelity, and temporal consistency of the video, respectively. We ask 20 subjects to rank different methods with 9 sets of comparisons in each study. From Tab. 1, the proposed zero-shot method achieves the best temporal consistency against baselines and shows a comparable frame-wise editing accuracy as the pre-frame optimization method [34]. As for the user studies, the average ranking of our method earns user preferences the best in three aspects.

To provide a qualitative comparison, Fig.7 provides the results of our method and other baselines at two different frames. The editing result of framewise SDEdit [32] can not be localized and varies a lot among different frames. Frame-wise Null inversion achieves local editing at the cost of 500-iterations optimization for each frame but is still temporally inconsistent. NLA-based [24] method preserves the exact pixels in the atlas. However, it struggles to perform editing that involves new shapes or 3D structures. In addition, it takes hours to optimize the neural atlas for each input video. While Tune-A-Video [51] with DDIM [45] ranks second in editing quality and image fidelity of Tab. 1, we observe that it has difficulty in reproducing the exact motion and spatial position as input video (right side of Fig.7). Besides, the background has annoying artifacts. Different from the above baselines, our method preserves the motion by fusion the attention during inversion and editing. Thus, our results outperform others by a large margin in our user study and frame consistency measured by CLIP.

4.4. Ablation Studies

Although we have proved the effectiveness of the proposed strategies in Fig. 4 and Fig. 3 using toy image examples, here, we ablate these designs in the video.

Attention during inversion. In the right column of Fig. 8, we use the attention map during reconstruction instead of inversion for zero-shot background editing. The visualized cross-attention map of the word `'boat'` in the first and last frame can not capture the correct position and structure of the boat, which may be caused by the poor temporal modeling capacity of the image diffusion model and the accumulation of errors in DDIM inversion. In contrast, we propose using attention during inversion as the middle column, which provides stable guidance of semantic layout in the original video. We observe this huge difference in attention maps between inversion and reconstruction exists in most videos.

Attention Blending Block is studied in Fig. 9, where we remove all self-attention fusion or fuse all self-attention without a spatial mask. The third column shows that removing all self-attention maps brings a loss of fine details (*e.g.*, fences, poles, and trees in the background) and inconsistency of car identity over time. In contrast, if we fuse full-resolution self-attention as in the previous work [16], the shape editing

ability of the framework can be severely degraded so that the geometry of generated car resembles the input video, especially in the last few frames. Therefore, we propose to blend the self-attention maps with a mask obtained from cross-attention to preserve unedited details and ensure temporal consistency while editing the object shape.

5. Conclusion

In this paper, we propose a new text-driven video editing framework `FateZero` that performs temporal consistent zero-shot editing of attribute, style, and shape. We make the first attempt to study and utilize the cross-attention and spatial-temporal self-attention during DDIM inversion, which provides fine-grained motion and structure guidance at each denoising step. A new Attention Blending Block is further proposed to enhance the shape editing performance of our framework. Our framework benefits **video** editing using widely existing **image** diffusion models, which we believe will contribute to a lot of new video applications.

Limitation & Future Work. While our method achieves impressive results, it still has some limitations. During shape editing, since the motion is produced by the one-shot video diffusion model [51], it is difficult to generate totally new motion (*e.g.*,'swim'→'fly') or very different shape (*e.g.*,'swan' →'pterosaur'). We will test our method on the generic pre-trained video diffusion model for better editing abilities.

References

[1] https://civitai.com, 2020. 2, 3

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 2, 3, 5

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 3, 5

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 2, 7

[5] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2015)*, 34(6), 2015. 2

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2, 3

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems*, 2021. 4

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 3

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3, 7, 8

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[11] Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sỳkora. Stylit: illumination-guided example-based stylization of 3d renderings. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[12] Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Sỳkora. Example-based synthesis of stylized facial animations. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017. 2

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 5

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 4, 5, 7, 8, 12

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *Empirical Methods in Natural Language Processing*, 2021. 8

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 4

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 5

[21] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Trans. Graph.*, 38(4), jul 2019. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference,*

Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 694–711. Springer, 2016. 2

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[24] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 2, 7, 8

[25] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2, 4

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[27] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 2

[28] Yao-Chih Lee, Ji-Ze Genevieve Jang Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing demo. *arXiv preprint arXiv:2301.13173*, 2023. 2, 3

[29] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020. 2

[30] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):356–371, 2022. 2

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. 2023. 7

[32] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3, 7, 8

[33] Alexander H. Miller, Will Feng, Dhruva Tirumala, Adam Fisch, Augustus Odena, Vivek Ramavajjala, Joel Z. Leibo, Kelvin Guu andJesse Engel, Jack Clark, Maruan H. Ali, Nazneen Rajani, Iain J. Dunning, Jacob Andreas, Chris Dyer, Dario Amodei, Jakob Uszkoreit, Douwe Pieksma, Tom Brown, and Ilya Sutskever. Clip: Learning to solve visual tasks by unsupervised learning of language representations. In *International Conference on Machine Learning*, 2020. 3, 8

[34] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 4, 5, 7, 8

[35] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2, 3

[36] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[37] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 2, 4, 8

[38] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2017. 7

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3, 4, 6, 7

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3

[44] 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. 2, 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 4, 7, 8

[46] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019. 4

[47] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 2, 3, 4

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[49] 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. 4, 5

[50] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain

textual descriptions. In *International Conference on Learning Representations*, 2023. 5

[51] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9

[52] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 2

[53] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[55] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3

A. Implementation Details

Pseudo algorithm code Our full algorithm is shown in Algorithm 1 and Algorithm 2. Algorithm 1 presents the overall framework of our inversion and editing, as visualized in the left of Fig. 1 in the main paper. Algorithm 2 shows that the cross-attention is fused based on a mask of the edited words, and the self-attention is blended using a binary mask from thresholding the cross-attention (the right of Fig. 1 in the main paper).

Hyperparameters Tuning. There are mainly three hyperparameters in our proposed designs:

- $t_s \in [1, T]$: Last timestep of the self-attention blending. Smaller t_s fuses more self-attention from inversion to preserve structure and motion.
- $t_c \in [1, T]$: Last timestep of the cross attention fusion. Smaller t_c fuses more cross attention from inversion to preserve the spatial semantic layout.
- $\tau \in [0, 1]$: Threshold for the blending mask used in shape editing. Smaller τ uses more self-attention map from editing to improve shape editing results.

In **style** and **attribute** editing, we set $t_s = 0.2T$, $t_c = 0.3T$, $\tau = 1.0$ to preserve most structure and motion in the source video. In **shape** editing, we set $t_s = 0.5T$, $t_c = 0.5T$, $\tau = 0.3$ to give more freedom in new motion and 3D shape generation.

B. Demo Video

we provide a detailed demo video to show:

Video Results on style, local attribute, and shape editing to validate the effectiveness of the proposed method.

Method Animation to provide a better understanding of the proposed method.

Baseline Comparisons with previous methods in video.

More Promising Applications We have shown the effectiveness of the proposed method in the main paper for style, attribution, and shape editing. In the demo video, we also show some potential applications of the proposed method, including (1) object removal by removing the word of the target object in the source prompt and mask the self-attention of the corresponding area using its cross attention, (2) video enhancement by adding the specific prompt (*e.g.*, 'high-quality', '8K') in the target editing prompt.

Algorithm 1 `FateZero` Algorithm

Input:
- z_0: Latent code from source video
- p_{src}: Source text prompt for input video
- p_{edit}: Target text prompt for edition

Hyperparameters:
- t_c: Last timestep of the cross attention fusion
- t_s: Last timestep of the self attention blending
- τ: Threshold for blending mask

Output:
- $\hat{z}_0$: Final edited latent code

▷ DDIM for inversion latents and attention maps
for $t = 1, 2, ..., T$ **do**
  $\epsilon_t, c_t^{\text{src}}, s_t^{\text{src}} \leftarrow \epsilon_\theta(z_t, t, p_{src})$
  $z_t = \sqrt{\alpha_t} \frac{z_{t-1} - \sqrt{1-\alpha_{t-1}}\epsilon_t}{\sqrt{\alpha_{t-1}}} + \sqrt{1-\alpha_t}\epsilon_t$
end for

▷ Denoising the inverted latents with attention fusion
for $t = T, (T-1), ..., 1$ **do**
  Edited_index $= (p_{src}$!= $p_{edit})$
  ▷ Cross-attention mask is from the edited index [16]
  M_{cross}[Edited_index] $= 1$
  ▷ Self-attention blending mask is from cross-attention.
  $M_{\text{self}} = (c_t^{\text{src}}$[Edited_index] $> \tau)$
  $\hat{\epsilon}_t \leftarrow$ ATT-FUSION$(\varepsilon_\theta, z_t, t, p_{\text{edit}}, M_{\text{edit}}, M_{\text{self}}, c_t^{\text{src}}, s_t^{\text{src}})$
  $z_{t-1} = \sqrt{\alpha_{t-1}} \frac{z_t - \sqrt{1-\alpha_t}\hat{\epsilon}_t}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\hat{\epsilon}_t$
end for
▷ Fuse the inversion and editing attention of all B blocks.
▷ We only show the operation of attention and omit the feed-forward, residual convolution layer for simplicity.
function ATT-FUSION$(\varepsilon_\theta, z_t, t, p_{\text{edit}}, M_{\text{cross}}, M_{\text{self}}, c_t^{\text{src}}, s_t^{\text{src}})$
  for $i = 1...B$ **do**
    $s_t^{\text{edit}} = \text{Softmax}(W_i^Q(z_t)W_i^K(z_t)/\sqrt{d_i})$
    $s_t^{\text{fused}} =$ SELF-BLENDING$(s_t^{\text{edit}}, s_t^{\text{src}}, M_{\text{self}}, c_t^{\text{src}}, t)$
    $z_t \quad = W_i^V(z_t) \cdot s_t^{\text{fused}}$
    $c_t^{\text{edit}} = \text{Softmax}(W_i^Q(z_t)W_i^K(p_{edit})/\sqrt{d_i})$
    $c_t^{\text{fused}} =$ CROSS-FUSION$(c_t^{\text{edit}}, c_t^{\text{src}}, M_{\text{edit}}, t)$
    $z_t \quad = W_i^V(p_{\text{edit}}) \cdot c_t^{\text{fused}}$
  end for
  return z_t
end function

C. Limitation and Future Work

Our zero-shot editing is not good at new concept composition or generation of very different shapes. For example, the result of editing 'black swan' to 'yellow pterosaur' in Fig 10 is unsatisfactory. This problem may be alleviated using a

Algorithm 2 Attention Fusion and Blending Algorithm

▷ Cross-attention fusion using the difference mask between source and editing prompt following prompt-to-prompt.
function CROSS-FUSION($c_t^{\text{edit}}, c_t^{\text{src}}, M_{\text{edit}}, t$)
  if $t > t_c$ **then**
    return $M_{\text{cross}} \cdot c_t^{\text{edit}} + (1 - M_{\text{cross}}) \cdot c_t^{\text{src}}$
  else
    return c_t^{edit}
  end if
end function

▷ Self-attention blending with cross attention.
function SLEF-BLENDING($s_t^{\text{edit}}, s_t^{\text{src}}, c_t^{\text{src}}, M_{\text{self}}, t$)
  if $t > t_s$ **then**
    return $M_{\text{self}} \cdot s_t^{\text{edit}} + (1 - M_{\text{self}}) \cdot s_t^{\text{src}}$
  else
    return s_t^{edit}
  end if
end function

black swan → yellow pterosaur.

Figure 10. limitation of our zero-shot editing.

stronger video diffusion model, which we leave to future work.