COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing

Jiangshan Wang[1]*, **Yue Ma**[1]*, **Jiayi Guo**[1]*, **Yicheng Xiao**[1], **Gao Huang**[1]†, **Xiu Li**[1]†
[1]Tsinghua University

https://cove-video.github.io/

Figure 1: We propose **CO**rrespondence-guided **V**ideo **E**diting (COVE), which leverages the correspondence information of the diffusion feature to achieve consistent and high-quality video editing. Our method is capable of generating high-quality edited videos with various kinds of prompts (style, category, background, etc.) while effectively preserving temporal consistency in generated videos.

Abstract

Video editing is an emerging task, in which most current methods adopt the pre-trained text-to-image (T2I) diffusion model to edit the source video in a zero-shot manner. Despite extensive efforts, maintaining the temporal consistency of edited videos remains challenging due to the lack of temporal constraints in the regular T2I diffusion model. To address this issue, we propose **CO**rrespondence-guided **V**ideo **E**diting (COVE), leveraging the inherent diffusion feature correspondence to achieve high-quality and consistent video editing. Specifically, we propose an efficient sliding-window-based strategy to calculate the similarity among tokens in the diffusion features of source videos, identifying the tokens with high correspondence across frames. During the inversion and denoising process, we

---

*Equal contribution. † Corresponding author.

sample the tokens in noisy latent based on the correspondence and then perform self-attention within them. To save GPU memory usage and accelerate the editing process, we further introduce the temporal-dimensional token merging strategy, which can effectively reduce redundancy. COVE can be seamlessly integrated into the pre-trained T2I diffusion model without the need for extra training or optimization. Extensive experiment results demonstrate that COVE achieves the start-of-the-art performance in various video editing scenarios, outperforming existing methods both quantitatively and qualitatively. The code will be release at https://github.com/wangjiangshan0725/COVE

1 Introduction

Diffusion models [24, 60, 62] have shown exceptional performance in image generation [54], thereby inspiring their application in the field of image editing [6, 22, 7, 50, 64, 21]. These approaches typically leverage a pre-trained Text-to-Image (T2I) stable diffusion model [54], using DDIM [61] inversion to transform source images into noise, which is then progressively denoised under the guidance of a prompt to generate the edited image.

Despite satisfactory performance in image editing, achieving high-quality video editing remains challenging. Specifically, unlike the well-established open-source T2I stable diffusion models [54], comparable T2V diffusion models are not as mature due to the difficulty of modeling complicated temporal motions, and training a T2V model from scratch demands substantial computational resources [23, 26, 59]. Consequently, there is a growing focus on adapting the pre-trained T2I diffusion for video editing [14, 30, 10, 69, 70, 51]. In this case, maintaining temporal consistency in edited videos is one of the biggest challenges, which requires the generated frames to be stylistically coherent and exhibit smooth temporal transitions, rather than appearing as a series of independent images. Numerous methods have been working on this topic while still facing various limitations, such as the inability to ensure fine-grained temporal consistency (leading to flickering [30, 51] or blurring [14] in generated videos), requiring additional components [27, 69, 10, 70] or needing extra training or optimization [70, 66, 38], etc.

In this work, our goal is to achieve highly consistent video editing by leveraging the intra-frame correspondence relationship among tokens, which is intuitively closely related to the temporal consistency of videos: If corresponding tokens across frames exhibit high similarity, the resulting video will thus demonstrate high temporal consistency. Taking a video of a man as an example, if the token representing his nose has high similarity across frames, his nose will be unlikely to deform or flicker throughout the video. However, how to obtain accurate correspondence information among tokens is still largely under-explored in existing works, although the intrinsic characteristic of the video editing task (i.e., the source video and edited

(a). Previous methods

• Based on pretrained Optical-flow model
• One-to-one correspondence

Frame i Frame $i+1$

(b). COVE

• Based on inherent diffusion feature correspondence
• One-to-many correspondence

Frame i Frame $i+1$

Figure 2: **Comparison between COVE (our method) and previous methods**[10, 70].

video are expected to share similar motion and semantic layout) determines that it naturally exists in the source video. Some previous methods [10, 70] leverage a pre-trained optical-flow model to obtain the flowing trajectory of each token across frames, which can be seen as a kind of coarse correspondence information. Despite the self-attention among tokens in the same trajectory can enhance the temporal consistency of the edited video, it still encounters two primary limitations: Firstly, these methods heavily rely on a highly accurate pre-trained optical-flow model to obtain the correspondence relationship of tokens, which is not available in many scenarios [29]. Secondly, supposing we have access to an extremely accurate optical-flow model, it is still only able to obtain the coarse one-to-one correspondence among tokens in different frames (Figure 2a), which would lead to the loss of information because one token is highly likely to correspond to multiple tokens in other frames in most cases (Figure 2b).
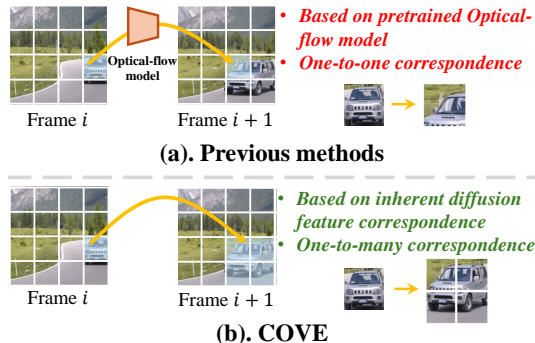
Addressing these problems, we notice that the inherent diffusion features naturally contain precise correspondence information. For instance, it is easy to find the corresponding points between two images by extracting their diffusion features and calculating the cosine similarity between tokens [63]. However, until now none of the existing works have successfully utilized this characteristic in more complicated and challenging tasks such as video editing. In this paper, we propose COVE, which is the first work unleashing the potential of inherent diffusion feature correspondence to significantly enhance the quality and temporal consistency in video editing. Given a source video, we first extract the diffusion feature of each frame. Then for each token in the diffusion feature, we obtain its corresponding tokens in other frames based on their similarity. Within this process, we propose a sliding-window-based approach to ensure computational efficiency. In our sliding-window-based method, for each token, it is only required to calculate the similarity between it and the tokens in the next frame located within a small window, identifying the tokens with the top K $(K > 1)$ highest similarity. After the correspondence calculation process, for each token, the coordinates of its K corresponding tokens in each other frame can be obtained. During the inversion and denoising process, we sample the tokens in noisy latents based on the obtained coordinates. To reduce the redundancy and accelerate the editing process, token merging is applied in the temporal dimension, which is followed by self-attention. Our method can be seamlessly integrated into the off-the-shelf T2I diffusion model without extra training or optimization. Extensive experiments demonstrate that COVE significantly improves both the quality and the temporal consistency of generated videos, outperforming a wide range of existing methods and achieving state-of-the-art results.

2 Related Works

2.1 Diffusion-based Image and Video Generation.

Diffusion Models [24, 60, 62] have recently showcased impressive results in image generation, which generates the image through gradual denoising from the standard Gaussian noise[11, 12, 49, 18, 54, 61, 19]. A large number of efforts on diffusion models [25, 31, 57] has enabled it to be applied to numerous scenarios [3, 13, 32, 36, 41, 46, 47, 55, 20, 44, 39]. With the aid of large-scale pretraining [52, 58], text-to-image diffusion models exhibit remarkable progress in generating diverse and high-quality images [48, 68, 53, 54, 56, 16, 45]. ControlNet [72] enables users to provide structure or layout information for precise generation. Naturally, diffusion models have found application in video synthesis, often by integrating temporal layers into image-based DMs [4, 23, 26, 67, 9]. Despite successes in unconditional video generation [26, 71, 42], text-to-video diffusion models lag behind their image counterparts.

2.2 Text-to-Video Editing.

There are increasing works adopting the pre-trained text-to-image diffusion model to the video editing task [40, 65, 66, 43, 17], where keeping the temporal consistency in the generated video is the most challenging. Recently, a large number of works focusing on zero-shot video editing has been proposed. FateZero [51] proposes to use attention blending to achieve high-quality edited videos while struggling to edit long videos. TokenFlow [14] reduces the effects of flickering through the linear combinations between diffusion features, while the smoothing strategy can cause blurring in the generated video. RAVE [30] proposes the randomized noise shuffling method, suffering the problem of fine details flickering. There are also a large number of methods that enhance the temporal consistency with the aid of pre-trained optical-flow models [70, 69, 10, 27]. Although the effectiveness of them, all of them severely rely on a pre-trained optical-flow model. Recent works [63] illustrate that the diffusion feature contains rich correspondence information. Although VideoSwap [15] adopts this characteristic by tracking the key points across frames, it still needs users to provide the key points as the extra addition manually.

3 Method

In this section, we will introduce COVE in detail, which can be seamlessly integrated into the pre-trained T2I diffusion model for high-quality and consistent video editing without the need for training or optimization (Figure 3). Specifically, given a source video, we first extract the diffusion feature of each frame using the pre-trained T2I diffusion model. Then, we calculate the one-to-many

(a). Sliding-window-based Correspondence Calculation

(b). Adopting the Correspondence in Attention

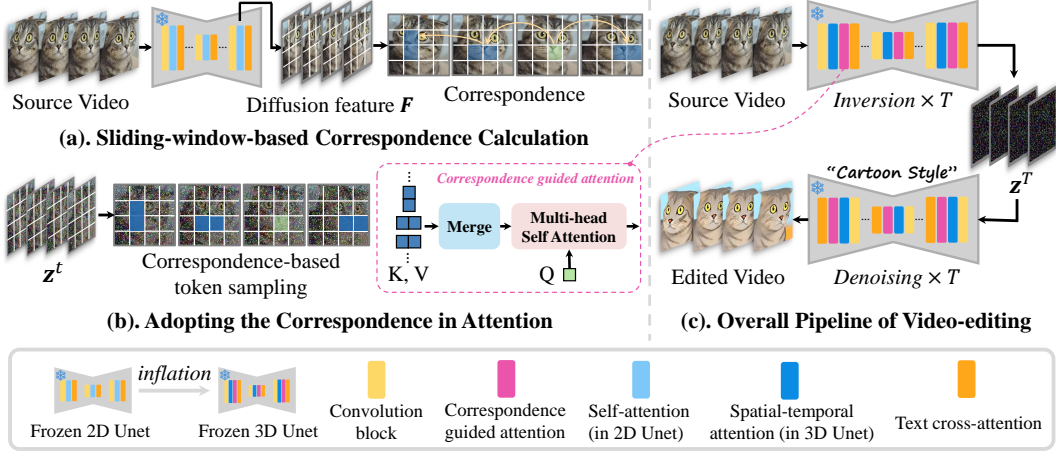**(c). Overall Pipeline of Video-editing**

Figure 3: **The overview of COVE.** (a). Given a source video, we extract the diffusion feature of each frame using the pre-trained T2I model and calculate the correspondence among tokens (detailed in Figure 4). (b). During the video editing process, we sample the tokens in noisy latent based on correspondence and apply self-attention among them. (c). The correspondence-guided attention can be seamlessly integrated into the T2I diffusion model for consistent and high-quality video editing.

correspondence of each token across frames based on cosine similarity (Figure 3a). To reduce resource consumption during correspondence calculation, we further introduce an efficient sliding-window-based strategy (Figure 4). During each timestep of inversion and denoising in video editing, the tokens in the noisy latent are sampled based on the correspondence and then merged. Through the self-attention among merged tokens (Figure 3b), the quality and temporal consistency of edited videos are significantly enhanced.

3.1 Preliminary

Diffusion Models. DDPM [24] is the latent generative model trained to reconstruct a fixed forward Markov chain $x_1, \ldots, x_T$. Given the data distribution $x_0 \sim q(x_0)$, the Markov transition $q(x_t|x_{t-1})$ is defined as a Gaussian distribution with a variance schedule $\beta_t \in (0, 1)$.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}). \tag{1}$$

To generate the Markov chain $x_0, \cdots, x_T$, DDPM leverages the reverse process with a prior distribution $p(x_T) = \mathcal{N}(x_T; 0, \mathbb{I})$ and Gaussian transitions. A neural network ϵ_θ is trained to predict noises, ensuring that the reverse process is close to the forward process.

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, \boldsymbol{\tau}, t), \Sigma_\theta(\boldsymbol{x}_t, \boldsymbol{\tau}, t)), \tag{2}$$

where $\boldsymbol{\tau}$ indicates the textual prompt. μ_θ and Σ_θ are predicted by the denoising model ϵ_θ. Since the diffusion and denoising process in the pixel space is computationally extensive, latent diffusion [54] is proposed to address this issue by performing these processes in the latent space of a VAE [34].

DDIM Inversion. DDIM can convert random noise to a deterministic $\boldsymbol{x}_0$ during sampling [61, 12]. The inversion process in deterministic DDIM can be formulated as follows:

$$\boldsymbol{x}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\boldsymbol{x}_t + \sqrt{\alpha_{t+1}}\left(\sqrt{\frac{1}{\alpha_{t+1}-1}} - \sqrt{\frac{1}{\alpha_t}-1}\right)\epsilon_\theta(\boldsymbol{x}_t), \tag{3}$$

where α_t denotes $\prod_{i=1}^t(1-\beta_i)$. The inversion process of DDIM is utilized to transform the input $\boldsymbol{x}_0$ into $\boldsymbol{x}_T$, facilitating subsequent tasks such as reconstruction and editing.

3.2 Correspondence Acquisition

As discussed in Section 1, intra-frame correspondence is crucial for the quality and temporal consistency of edited videos while remaining largely under-explored in existing works. In this section, we introduce our method for obtaining correspondence relationships among tokens across frames.
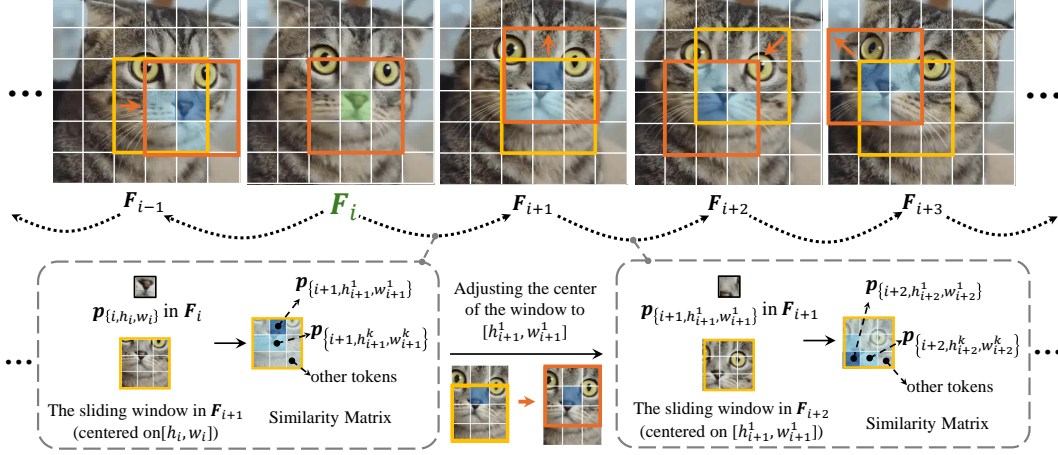
4

Figure 4: **Sliding-window-based strategy for correspondence calculation.** ▢ represents the token $\boldsymbol{p}_{\{i,h_i,w_i\}}$. ▢ and ▢ represents the obtained corresponded tokens in other frames.

Diffusion Feature Extraction. Given a source video $\boldsymbol{V}$ with N frames, a VAE [34] is employed on each frame to extract the latent features $\boldsymbol{Z} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$, where $\boldsymbol{Z} \in \mathbb{R}^{N \times H \times W \times d}$. Here, H and W denote the height and width of the latent feature and d denotes the dimension of each token. For each frame of $\boldsymbol{Z}$, we add noise of a specific timestep t and feed the noisy frame $\boldsymbol{Z}^t = \{\boldsymbol{z}_1^t, \cdots, \boldsymbol{z}_N^t\}$ into the pre-trained T2I model f_θ respectively. The diffusion feature (i.e., the intermediate feature from the U-Net decoder) is extracted through a single step of denoising [63]:

$$\boldsymbol{F} = \{\boldsymbol{F}_i\} = \{f_\theta(\boldsymbol{z}_i^t)\}, i \in \{1, \cdots, N\}, \tag{4}$$

where $\boldsymbol{F} \in \mathbb{R}^{N \times H \times W \times d}$, denoting the normalized diffusion feature of each frame.

One-to-many Correspondence Calculation. For each token within the diffusion feature $\boldsymbol{F}$, its corresponding tokens in other frames are identified based on the cosine similarity. Without loss of generality, we could consider a specific token $\boldsymbol{p}_{\{i,h_i,w_i\}}$ in the ith frame $\boldsymbol{F}_i$ with the coordinate $[h_i, w_i]$. Unlike previous methods where only one corresponding token of $\boldsymbol{p}_{\{i,h_i,w_i\}}$ can be identified in each frame (Figure 2a), our method can obtain the one-to-many correspondences simply by selecting tokens with the top K highest similarity in each frame. We record their coordinates, which are used for sampling the tokens for self-attention in the subsequent inversion and denoising process. To implement this process, the most straightforward method is through a direct matrix multiplication of the normalized diffusion feature $\boldsymbol{F}$.

$$\boldsymbol{S} = \boldsymbol{F} \cdot \boldsymbol{F}^T, \tag{5}$$

where $\boldsymbol{S} \in \mathbb{R}^{(N \times H \times W) \times (N \times H \times W)}$ represents the cosine similarity between each token and all tokens in the diffusion feature of the video.

The similarity between $\boldsymbol{p}_{\{i,h_i,w_i\}}$ and all $N \times H \times W$ tokens in the feature is given by $\boldsymbol{S}[i, h_i, w_i, :, :, :]$. The coordinates of the corresponding tokens in the jth frame ($j \in \{1, \cdots, N\}$) are then obtained by selecting the tokens with the top K similarities in the jth frame.

$$h_j^k, w_j^k = \text{top-}k\text{-argmax}_{(x^k, y^k)}(\boldsymbol{S}[i, h_i, w_i, j, x^k, y^k]), \tag{6}$$

Here the top-k-argmax($\cdot$) denotes the operation to find coordinates of the top K biggest values in a matrix, where $k \in \{1, \cdots, K\}$. $[h_j^k, w_j^k]$ represents the coordinates of the token in jth frame which has highest similarity with $\boldsymbol{p}_{\{i,h_i,w_i\}}$. A similar process can be conducted for each token of $\boldsymbol{F}$, thereby obtaining their correspondences among frames.

Sliding-window Strategy. Although the one-to-many correspondence among tokens can be effectively obtained through the above process, it requires excessive computational resources because $(N \times H \times W)$ is always a huge number, especially in long videos. As a result, the computational complexity of this process is extremely high, which can be represented as $\mathcal{O}(N^2 \times H^2 \times W^2 \times d)$. At the same time, multiplication between these two huge matrices consumes a substantial amount

of GPU memory in practice. These limitations severely limit its applicability in many real-world scenarios, such as on mobile devices.

To address the above problem, we further propose the sliding-window-based strategy as an alternative, which not only effectively obtains the one-to-many correspondences but also significantly reduces the computational overhead (Figure 4). Firstly, for the token $\boldsymbol{p}_{\{i,h_i,w_i\}}$, it is only necessary to calculate its similarity with the tokens in the next frame $\boldsymbol{F}_{i+1}$ instead of in all frames, i.e.,

$$\boldsymbol{S}_i = \boldsymbol{F}_i \cdot \boldsymbol{F}_{i+1}^T. \tag{7}$$

$\boldsymbol{S}_i \in \mathbb{R}^{H \times W \times H \times W}$ denotes the similarity between the tokens in ith frame and those in $(i+1)$th frame. The overall similarity matrix is $\boldsymbol{S} = \{\boldsymbol{S}_i\}, i \in \{1, 2, \cdots, N-1\}$, where $\boldsymbol{S} \in \mathbb{R}^{(N-1) \times H \times W \times H \times W}$. Then, we obtain the K corresponded tokens of $\boldsymbol{p}_{\{i,h_i,w_i\}}$ in $\boldsymbol{F}_{i+1}$ through $\boldsymbol{S}_i$,

$$h_{i+1}^k, w_{i+1}^k = \text{top-}k\text{-argmax}_{(x^k, y^k)}(\boldsymbol{S}_i[h_i, w_i, x^k, y^k]), \tag{8}$$

For tokens in $(i + 2)$th frame, instead of considering $\boldsymbol{p}_{\{i,h_i,w_i\}}$, we identify the tokens in $(i + 2)$th frame which have the top K largest similarity with the token $\boldsymbol{p}_{\{i+1,h_{i+1}^1,w_{i+1}^1\}}$ through the $\boldsymbol{S}_{i+1}$. Similarly, we can obtain the corresponding token in other future or previous frames.

$$h_{i+2}^k, w_{i+2}^k = \text{top-}k\text{-argmax}_{(x^k, y^k)}(\boldsymbol{S}_{i+1}[h_{i+1}^1, w_{i+1}^1, x^k, y^k]), \tag{9}$$

Through the above process, the overall complexity is reduced to $\mathcal{O}((N-1) \times H^2 \times W^2 \times d)$. Furthermore, it is noteworthy that frames in a video exhibit temporal continuity, implying that the spatial positions of corresponding tokens are unlikely to change significantly between consecutive frames. Consequently, for the token $\boldsymbol{p}_{\{i,h_i,w_i\}}$, it is enough to only calculate the similarity within a small window of length l in the adjacent frame, where l is much smaller than H and W,

$$\boldsymbol{F}_{i+1}^w = \boldsymbol{F}_{i+1}[h_i - l/2 : h_i + l/2, w_i - l/2 : w_i + l/2, :]. \tag{10}$$

$\boldsymbol{F}_{i+1}^w \in \mathbb{R}^{l \times l \times d}$ represents the tokens in $\boldsymbol{F}_{i+1}$ within the sliding window. We calculate the cosine similarity between $\boldsymbol{p}_{\{i,h_i,w_i\}}$ and the tokens in $\boldsymbol{F}_{i+1}^w$, selecting tokens with top K highest similarity within $\boldsymbol{F}_{i+1}^w$. This approach further reduces the computational complexity to $\mathcal{O}((N-1) \times H \times W \times l^2 \times d)$ and the GPU memory consumption is also significantly reduced in practice. Additionally, it is worth noting that calculating correspondence information from the source video is only conducted once before the inversion and denoising process of video editing. Compared with the subsequent editing process, this process only takes negligible time.

3.3 Correspondence-guided Video Editing.

In this section, we explain how to apply the correspondence information to the video editing process (Figure 3c). In the inversion and denoising process of video editing, we sample the corresponding tokens from the noisy latent for each token based on the coordinates obtained in Section 3.2. For the token $\boldsymbol{z}_{i,h_i,w_i}^t$, the set of corresponding tokens in other frames at a timestep t is:

$$\boldsymbol{Corr} = \{\boldsymbol{z}_{\{j,h_j^k,w_j^k\}}^t\}, j \in \{1, \cdots, i-1, i+1, \cdots, N\}, k \in \{1, \cdots, K\}. \tag{11}$$

We merge these tokens following [5], which can accelerate the editing process and reduce GPU memory usage without compromising the quality of editing results:

$$\widetilde{\boldsymbol{Corr}} = \text{Merge}(\boldsymbol{Corr}). \tag{12}$$

Then, the self-attention is conducted on the merged tokens,

$$\boldsymbol{Q} = \boldsymbol{z}_{\{i,h_i,w_i\}}^t, \boldsymbol{K} = \boldsymbol{V} = \widetilde{\boldsymbol{Corr}}, \tag{13}$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{SoftMax}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{d_k}}\right) \cdot \boldsymbol{V}, \tag{14}$$

where $\sqrt{d_k}$ is the scale factor. The above process of correspondence-guided attention is illustrated in Figure 3b. Following the previous methods [70, 10], we also retain the spatial-temporal attention [66] in the U-Net. In spatial-temporal attention, considering a query token, all tokens in the video serve as keys and values, regardless of their relevance to the query. This correspondence-agnostic self-attention is not enough to maintain temporal consistency, introducing irrelevant information into each token, and thus causing serious flickering effects [10, 14]. Our correspondence-guided attention can significantly alleviate the problems of spatial-temporal attention, increasing the similarity of corresponding tokens and thus enhancing the temporal consistency of the edited video.
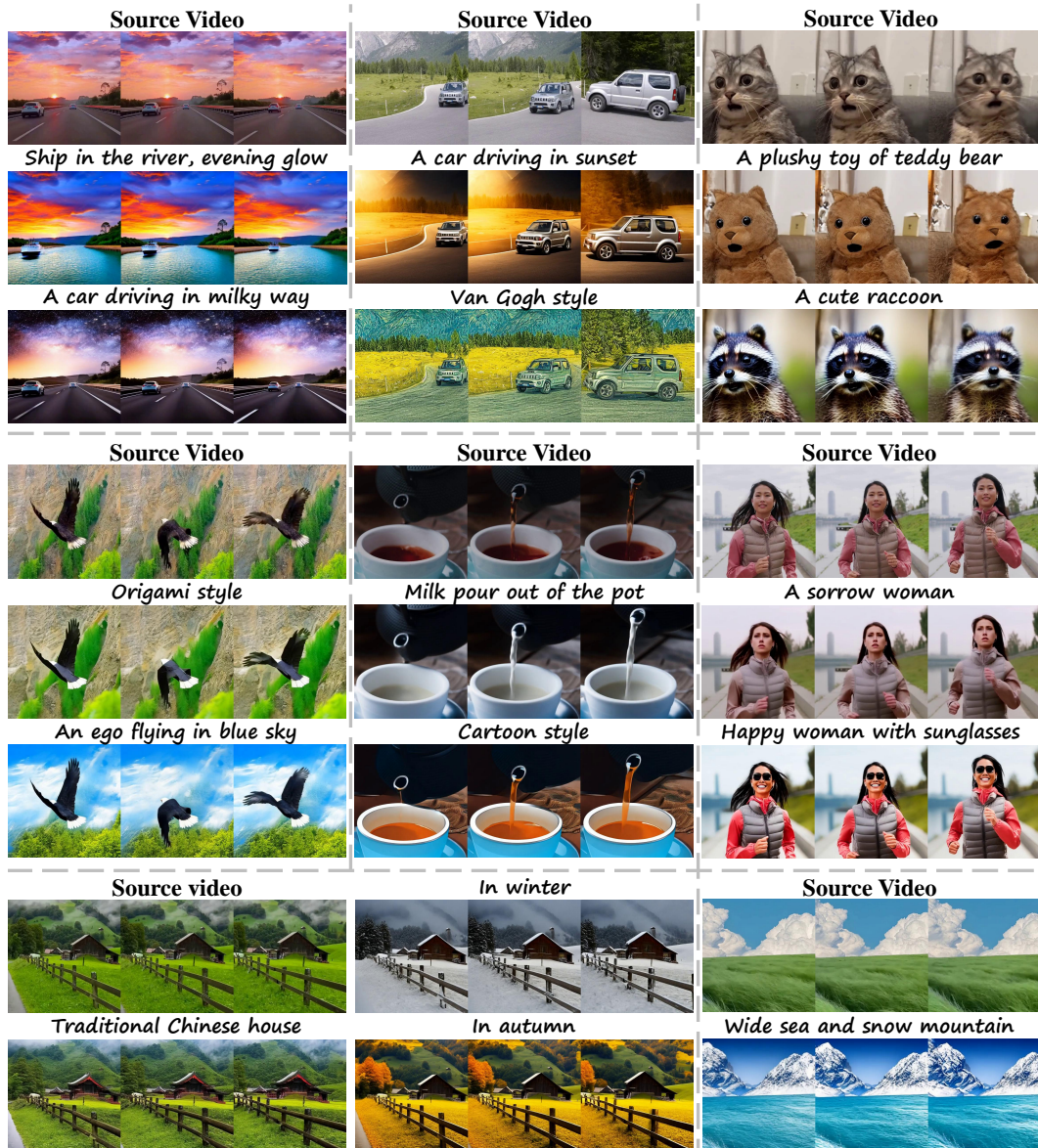
6

Figure 5: **Qualitative results of COVE.** COVE can effectively handle various types of prompts, generating high-quality videos. For both global editing (e.g., style transferring and background editing) and local editing (e.g., modifying the appearance of the subject), COVE demonstrates outstanding performance. Results are best-viewed zoomed-in.

4 Experiment

4.1 Experimental Setup

In the experiment, we adopt Stable Diffusion (SD) 2.1 from the official Huggingface repository for COVE, employing 100 steps of DDIM inversion and 50 steps of denoising. To extract the diffusion feature, the noise of the specific timestep $t = 261$ is added to each frame of the source video following [63]. The feature is then extracted from the intermediate layer of the 2D Unet decoder during a single step of denoising. The window size l is set to 9 for correspondence calculation, and k is set to 3 for correspondence-guided attention. The merge ratio for token merging is 50%. For both qualitative and quantitative evaluation, we select 23 videos from social media platforms such as TikTok and other publicly available sources [1, 2]. Among these 23 videos, 3 videos have a length of 10 frames, 15 videos have a length of 20 frames, and 5 videos have a length of 32 frames. The experiments are
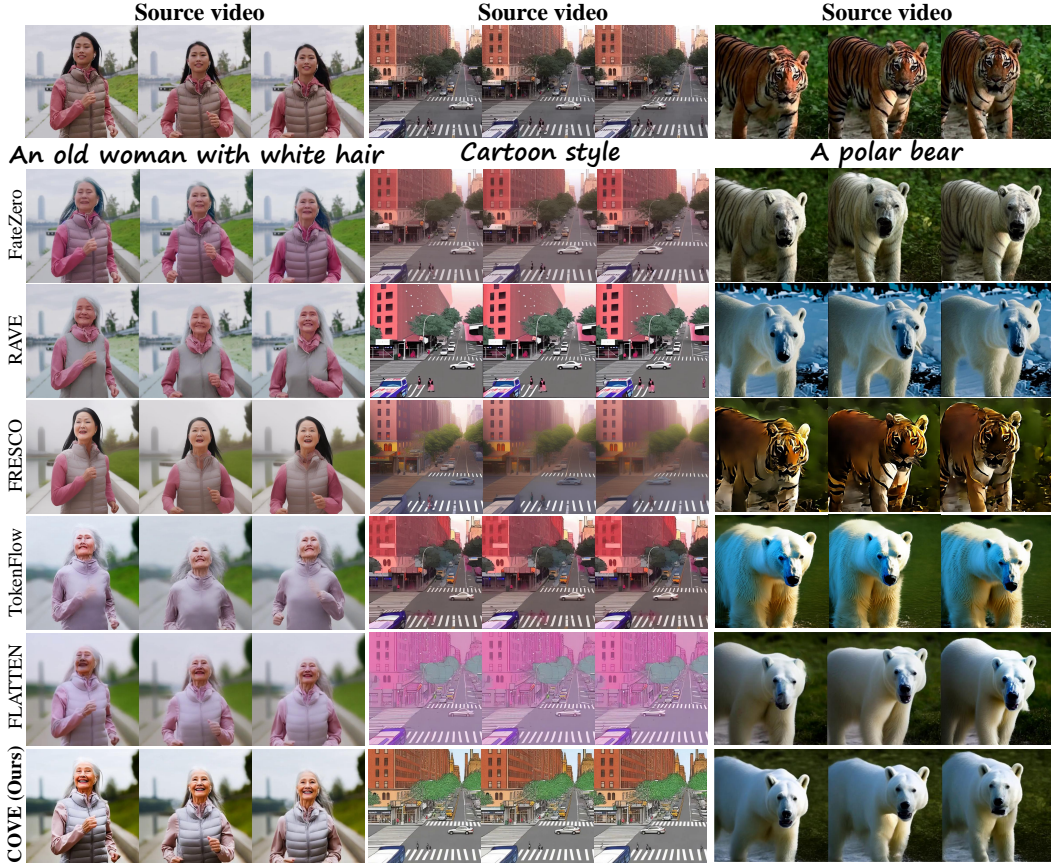
Figure 6: **Qualitative comparison of COVE and various state-of-the-art methods.** Our method outperforms previous methods across a wide range of source videos and editing prompts, demonstrating superior visual quality and temporal consistency. Results are best-viewed zoomed-in.

conducted on a single RTX 3090 GPU for our method unless otherwise specified. We compare COVE with 5 baseline methods: FateZero [51], TokenFlow [14], FLATTEN [10], FRESCO [70] and RAVE [30]. For all of these baseline methods, we follow the default settings from their official Github repositories. The more detailed experimental settings of our method are provided in Appendix A.

4.2 Qualitative Results

We evaluate COVE on various videos under different types of prompts including both global and local editing (Figure 5). Global editing mainly involves background editing and style transferring. For background editing, COVE can modify the background while keeping the subject of the video unchanged (e.g. Third row, first column. "`a car driving in milky way`"). For style transfer, COVE can effectively modify the global style of the source video according to the prompt (e.g. Third row, second column. "`Van Gogh style`"). Our prompts for local editing include changing the subject of the video to another one (e.g. Third row, third column. "`A cute raccoon`") and making local edits to the subject (e.g. fifth row, third column. "`A sorrow woman`"). For all of these editing tasks, COVE demonstrates outstanding performance, generating frames with high visual quality while successfully preserving temporal consistency. We also compare COVE with a wide range of state-of-the-art video editing methods (Figure 6). The experimental results illustrate that COVE effectively edits the video with high quality, significantly outperforming the previous methods.

4.3 Quantitative Results

For quantitative comparison, we follow the metrics proposed in VBench [28], including Subject Consistency, Motion Smoothness, Aesthetic Quality, and Imaging Quality. Among them, Subject Consistency assesses whether the subject (e.g., a person) remains consistent throughout the whole

8

video by calculating the similarity of DINO [8] feature across frames. Motion Smoothness utilizes the motion priors of the video frame interpolation model [37] to evaluate the smoothness of the motion in the generated video. Aesthetic Quality uses the LAION aesthetic predictor [35] to assess the artistic and beauty value perceived by humans on each frame. Imaging Quality evaluates the degree of distortion in the generated frames (e.g., blurring, flickering) through the MUSIQ [33] image quality predictor. Each video undergoes editing with 3 global prompts (such as style transferring, background editing, etc.) and 2 local prompts (such as editing the appearance of the subject in the video), generating a total of 115 text-video pairs. For each metric, we report the average score of these 115 videos. We further conducted a user study with 45 participants following [70]. Participants are required to choose the most preferable results among these methods. The result is shown in Table 1. Among various methods, COVE achieves outstanding performance in both qualitative metrics and user studies, further demonstrating its superiority.

	Subject Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	User Study
FateZero [51]	0.9622	0.9547	0.6258	0.6951	7.4%
TokenFlow [14]	0.9513	0.9803	0.6904	0.7354	13.0%
FLATTEN [10]	0.9617	0.9622	0.6544	0.7155	14.8%
FRESCO [70]	0.9358	0.9737	0.6582	0.6331	9.2%
RAVE [30]	0.9518	0.9732	0.6369	0.7355	11.1%
COVE (Ours)	**0.9731**	**0.9892**	**0.7122**	**0.7441**	**44.5%**

Table 1: **Quantitative comparison** among COVE and a wide range of state-of-the-art video editing methods. The evaluation metrics[28] can effectively reflect the temporal consistency and frame quality of generated videos. COVE illustrates superior performance in both keeping the temporal consistency and generating frames with high quality in edited videos.

	Subject Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality
w/o	0.9431	0.9049	0.6913	0.7132
$K = 1$	0.9637	0.9817	0.6979	0.7148
$K = 3$	0.9731	**0.9892**	0.7122	**0.7441**
$K = 5$	**0.9745**	0.9886	**0.7167**	0.7429

Table 2: **Ablation study on the value of K in correspondence-guided attention.** w / o means without correspondence-guided attention in Unet. When $K = 3$ the quality of the video is the best.
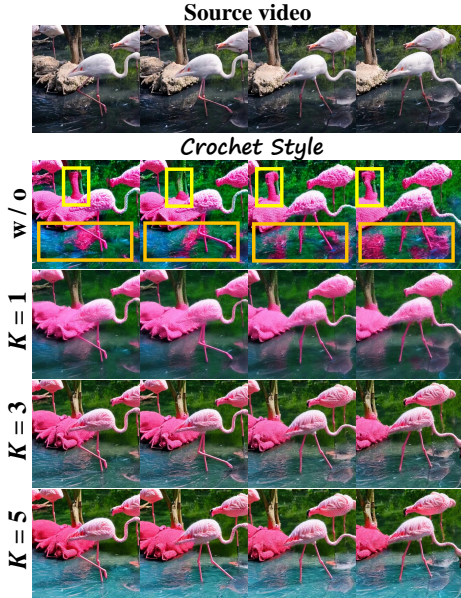

Figure 7: Ablation study about the correspondence-guided attention and the value of K. w / o means do not apply correspondence-guided attention.

4.4 Ablation Study

We conduct an ablation study to illustrate the effectiveness of the **Correspondence-guided attention** and the number of tokens selected in each frame (i.e., the value of K). The experimental results (Table 2 and Figure 7) illustrate that without correspondence-guided attention, the edited video exhibits obvious temporal inconsistency and flickering effects (which is marked in yellow and orange boxes in Figure 7), thus severely impairing the visual quality. As K increases from 1 to 3, the generated video contains more fine-grained details, exhibiting better visual quality. However, further increasing K to 5 does not significantly improve the video quality. We also illustrate the effectiveness of **temporal dimensional token merging**. By merging the tokens with high correspondence across frames, the editing process becomes more efficient (Table 3) while there is no significant decrease in the quality of the edited video (Figure 8). The ablation of the **sliding-window size** l is shown in Appendix B. If the window size is too small, the actual corresponding token may not be included within the window, resulting in suboptimal correspondence and poor editing results. On the other hand, a too-large window size is not necessary for identifying the corresponding tokens, which would lead to high computational complexity and excessive memory usage. The experiment results illustrate that $l = 9$ is suitable to strike a balance. Additionally, we also **visualize the correspondence** obtained by COVE, which is shown in Appendix C.

Correspondence Guided Attention	Token Merging	Speed	Memory Usage
✗	✗	2.2 min	9 GB
✓	✗	2.7 min	14 GB
✓	✓	2.4 min	11 GB

Table 3: **Ablation Study on the effect of temporal dimensional token merging.** Temporal dimensional token merging can speed up the editing process and save GPU memory usage while hardly impairing the quality of the generated video. The experiment is conducted on a single RTX3090 GPU with a 10-frame source video. k is set to 3.

Figure 8: Token merging would not impair the quality of edited video.

5 Conclusion

In this paper, we propose COVE, which is the first to explore how to employ inherent diffusion feature correspondence in video editing to enhance editing quality and temporal consistency. Through the proposed efficient sliding-window-based strategy, the one-to-many correspondence relationship among tokens across frames is obtained. During the inversion and denoising process, self-attention is performed within the corresponding tokens to enhance temporal consistency. Additionally, we also apply token merging in the temporal dimension to improve the efficiency of the editing process. Both quantitative and qualitative experimental results demonstrate the effectiveness of our method, which outperforms a wide range of previous methods, achieving state-of-the-art editing quality.

Limitaions. The limitation of our method is discussed in Appendix E.

References

[1] Pexels. https://www.pexels.com/, accessed: 2023-11-16 7

[2] Pixabay. https://pixabay.com/, accessed: 2023-11-16 7

[3] Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022) 3

[4] 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023) 3

[5] Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022) 6

[6] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 2

[7] Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22560–22570 (2023) 2

[8] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 9

[9] Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: The Twelfth International Conference on Learning Representations (2023) 3

[10] Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: Flatten: optical flow-guided attention for consistent text-to-video editing. arXiv preprint arXiv:2310.05922 (2023) 2, 3, 6, 8, 9

[11] Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3

[12] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 3, 4

[13] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 3

[14] Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023) 2, 3, 6, 8, 9

[15] Gu, Y., Zhou, Y., Wu, B., Yu, L., Liu, J.W., Zhao, R., Wu, J.Z., Zhang, D.J., Shou, M.Z., Tang, K.: Videoswap: Customized video subject swapping with interactive semantic point correspondence. arXiv preprint arXiv:2312.02087 (2023) 3

[16] Guo, J., Du, C., Wang, J., Huang, H., Wan, P., Huang, G.: Assessing a single image in reference-guided image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 753–761 (2022) 3

[17] Guo, J., Manukyan, H., Yang, C., Wang, C., Khachatryan, L., Navasardyan, S., Song, S., Shi, H., Huang, G.: Faceclip: Facial image-to-video translation via a brief text description. IEEE Transactions on Circuits and Systems for Video Technology (2023) 3

[18] Guo, J., Wang, C., Wu, Y., Zhang, E., Wang, K., Xu, X., Song, S., Shi, H., Huang, G.: Zero-shot generative model adaptation via image-specific prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11494–11503 (2023) 3

[19] Guo, J., Xu, X., Pu, Y., Ni, Z., Wang, C., Vasu, M., Song, S., Huang, G., Shi, H.: Smooth diffusion: Crafting smooth latent spaces in diffusion models. arXiv preprint arXiv:2312.04410 (2023) 3

[20] Guo, J., Zhao, J., Ge, C., Du, C., Ni, Z., Song, S., Shi, H., Huang, G.: Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. arXiv preprint arXiv:2406.04295 (2024) 3

[21] Guo, Q., Lin, T.: Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. arXiv preprint arXiv:2312.10113 (2023) 2

[22] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 2

[23] 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) 2, 3

[24] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020) 2, 3, 4

[25] Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 3

[26] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022) 2, 3

[27] Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. arXiv preprint arXiv:2307.14073 (2023) 2, 3

[28] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. arXiv preprint arXiv:2311.17982 (2023) 8, 9

[29] Jonschkowski, R., Stone, A., Barron, J.T., Gordon, A., Konolige, K., Angelova, A.: What matters in unsupervised optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 557–572. Springer (2020) 2

[30] Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. arXiv preprint arXiv:2312.04524 (2023) 2, 3, 8, 9

[31] Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems 35, 26565–26577 (2022) 3

[32] Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems 35, 23593–23606 (2022) 3

[33] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021) 9

[34] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 4, 5

[35] LAION-AI: aesthetic-predictor. https://github.com/LAION-AI/aesthetic-predictor (2022) 9

[36] Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing 479, 47–59 (2022) 3

[37] Li, Z., Zhu, Z.L., Han, L.H., Hou, Q., Guo, C.L., Cheng, M.M.: Amt: All-pairs multi-field transforms for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9801–9810 (2023) 9

[38] Liew, J.H., Yan, H., Zhang, J., Xu, Z., Feng, J.: Magicedit: High-fidelity and temporally coherent video editing. arXiv preprint arXiv:2308.14749 (2023) 2

[39] Lin, Y., Han, H., Gong, C., Xu, Z., Zhang, Y., Li, X.: Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. arXiv preprint arXiv:2309.17261 (2023) 3

[40] Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023) 3

[41] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2022) 3

[42] Ma, Y., Cun, X., He, Y., Qi, C., Wang, X., Shan, Y., Li, X., Chen, Q.: Magicstick: Controllable video editing via control handle transformations. arXiv preprint arXiv:2312.03047 (2023) 3

[43] Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024) 3

[44] Ma, Y., He, Y., Wang, H., Wang, A., Qi, C., Cai, C., Li, X., Li, Z., Shum, H.Y., Liu, W., et al.: Follow-your-click: Open-domain regional image animation via short prompts. arXiv preprint arXiv:2403.08268 (2024) 3

[45] Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. arXiv preprint arXiv:2406.01900 (2024) 3

[46] 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. arXiv preprint arXiv:2108.01073 (2021) 3

[47] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023) 3

[48] Nichol, A., 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. arXiv preprint arXiv:2112.10741 (2021) 3

[49] Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021) 3

[50] Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 2

[51] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023) 2, 3, 8, 9

[52] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3

[53] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 3

[54] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3, 4

[55] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 3

[56] 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. Advances in neural information processing systems **35**, 36479–36494 (2022) 3

[57] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022) 3

[58] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 3

[59] 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. arXiv preprint arXiv:2209.14792 (2022) 2

[60] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 2, 3

[61] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 2, 3, 4

[62] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) 2, 3

[63] Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023) 3, 5, 7, 15

[64] Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023) 2

[65] Wang, W., Jiang, Y., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint arXiv:2303.17599 (2023) 3

[66] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023) 2, 3, 6

[67] Xiang, J., Huang, R., Zhang, J., Li, G., Han, X., Wei, Y.: Versvideo: Leveraging enhanced temporal diffusion models for versatile video generation. In: The Twelfth International Conference on Learning Representations (2023) 3

[68] Xue, J., Wang, H., Tian, Q., Ma, Y., Wang, A., Zhao, Z., Min, S., Zhao, W., Zhang, K., Shum, H.Y., et al.: Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. arXiv e-prints pp. arXiv–2406 (2024) 3

[69] Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023) 2, 3

[70] Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Fresco: Spatial-temporal correspondence for zero-shot video translation. arXiv preprint arXiv:2403.12962 (2024) 2, 3, 6, 8, 9

[71] Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18456–18466 (2023) 3

[72] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 3

Appendix

A Detailed Experimental Settings

In the experiment, the size of all source videos is 512×512. We adopt Stable Diffusion (SD) 2.1 from the official Huggingface repository for our method. To extract the diffusion feature, following [63], the noise of the timestep $t = 261$ is added to each frame of the source video. The noisy frames of video are fed into the U-net, the feature is extracted from the intermediate layer of the 2D Unet decoder. The height and weight of the diffusion feature is 64. Following previous works, at the first 40 timesteps, the diffusion features are saved during DDIM inversion and are further injected during denoising. For Spatial-temporal attention, we use the xFormers to reduce memory consumption, while it is not used in correspondence-guided attention.

B Ablation Study on the Window Size

To illustrate the influence of window size l, we conduct the experiment on a video with 20 frames on a single A100 GPU. During the correspondence calculation process, we calculate the theoretical computational complexity, which is the total number of multiplications and additions required. We also record actual GPU memory consumed under different window sizes, the result is shown in Appendix B. With our sliding window strategy, the computational complexity and the GPU memory in the correspondence calculation process are significantly reduced. The visualization result is shown in Fig. 9. If the window size is too small, the motion in the video cannot be tracked, causing unsatisfying results. We choose $l = 9$ for the experiments in other sections, which can achieve a balance between the memory consumed and the quality of the edited video.

Window Size (l)	3	9	15	w/o
Computational Complexity ($\times 10^9$)	0.448	4.03	11.2	241
GPU Memory (GB)	11	14	18	32

Table 4: **Ablation Study on the window size** l. w/o means that the sliding-window strategy is not applied. The sliding window strategy can significantly reduce the use of computational complexity and GPU memory.
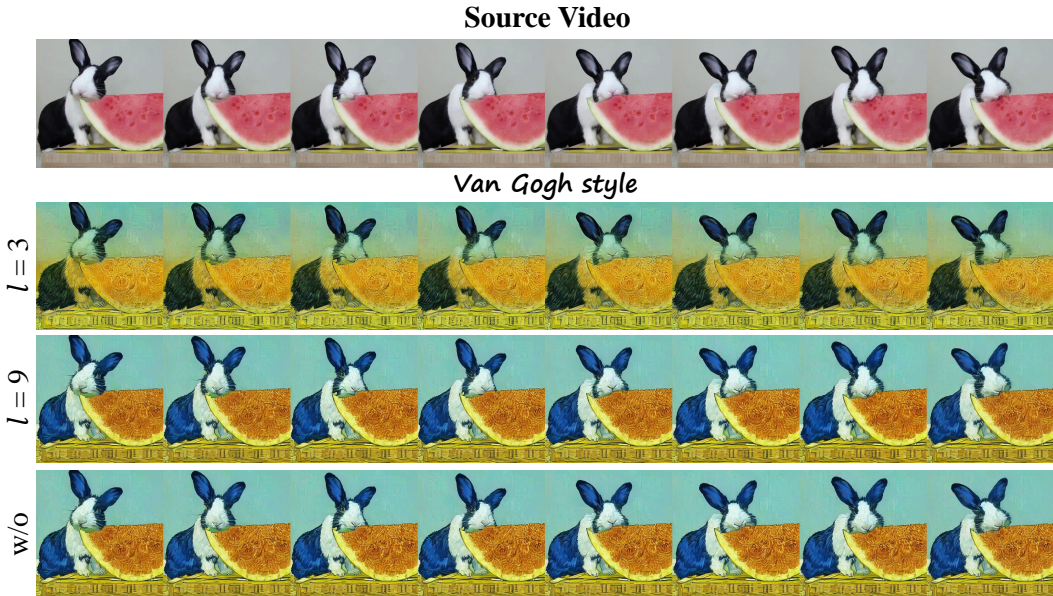

Figure 9: **Ablation Study on the window size** l.

C Visualisation of the Correspondence

We visualize the correspondence calculated by our sliding-window-based method to illustrate its effectiveness (Fig. 10). To be specific, we calculate the correspondence based on the 64×64 diffusion feature, which is extracted at the final layer of the U-net decoder. We select the token representing the left eye of the wolf in the 7th frame (marked in red) and visualize its corresponding tokens in other frames (marked in yellow). The result illustrates that our method can effectively identify the corresponding tokens.

Figure 10: **Visualsation of the correspondence calculated by our method.**

D Broader Impacts

Our work enables high-quality video editing, which is in high demand across various social media platforms, especially short video websites like TikTok. Using our method, people can easily create high-quality and creative videos, significantly reducing production costs. However, there is a potential for misuse, such as replacing the characters in videos with celebrities, which may infringe upon the celebrities' image rights. Therefore, it is also necessary to improve relevant laws and regulations to ensure the legal use of our method.

E Limitations

Despite achieving outstanding results, our methods still encounter several limitations. First, although the correspondence calculation process is efficient through the proposed sliding window strategy, the implementation of correspondence-guided attention is still not efficient enough, leading to the extra usage of GPU memory and time (Table 3). This problem is expected to be alleviated largely through the use of xFormers. We will work on it in the future.

Second, further exploration is required to optimize the application of the obtained correspondence information. In this study, we utilize the correspondence information to sample tokens during the inversion and denoising processes and do the self-attention. However, we believe that there may be more effective alternatives to self-attention that could further unleash the potential of the correspondence information.

F More Qualitative Results

We provide more qualitative results of our method to illustrate its effectiveness, which is shown in Fig. 12 and Fig. 11.

16

Source Video

A red car in snowy winter

A Wooden car

A car driving in dry desert

A car, oil painting style

A car driving in sunset

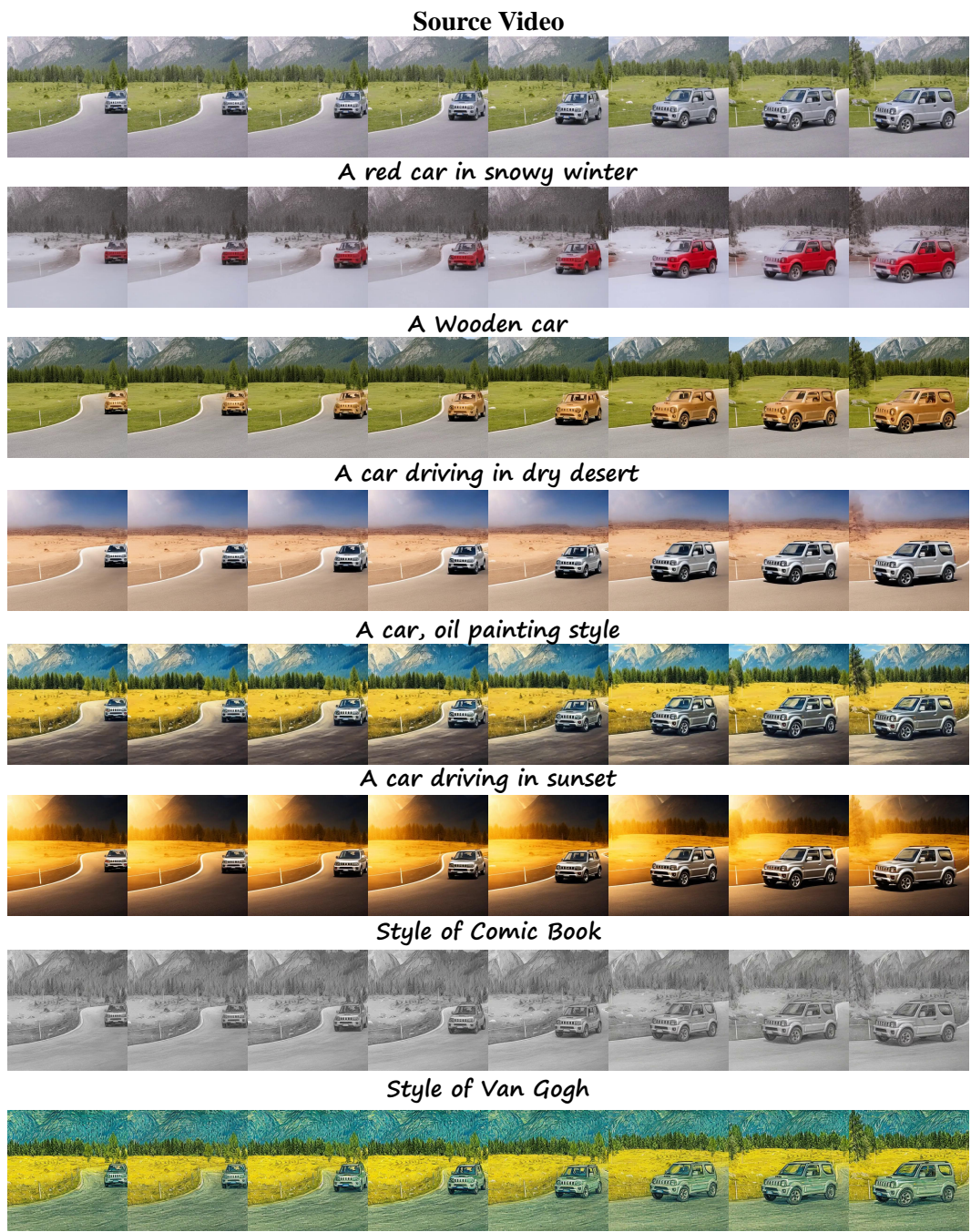
Style of Comic Book

Style of Van Gogh

Figure 11: **Qualitative results of our methods.** Our method can effectively handle various kinds of prompts, generating high-quality videos. Results are best viewed in zoomed-in.

Source Video

An African woman in grey clothes

A woman dressed like Black widow

A marble sculpture

Cartoon style

Figure 12: **Qualitative results of our methods.** Our method can effectively handle various kinds of prompts, generating high-quality videos. Results are best viewed in zoomed-in.