COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing

Jiangshan Wang[1*], **Yue Ma**[2*], **Jiayi Guo**[1*], **Yicheng Xiao**[1], **Gao Huang**[1†], **Xiu Li**[1†]
[1]Tsinghua University, [2]HKUST

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

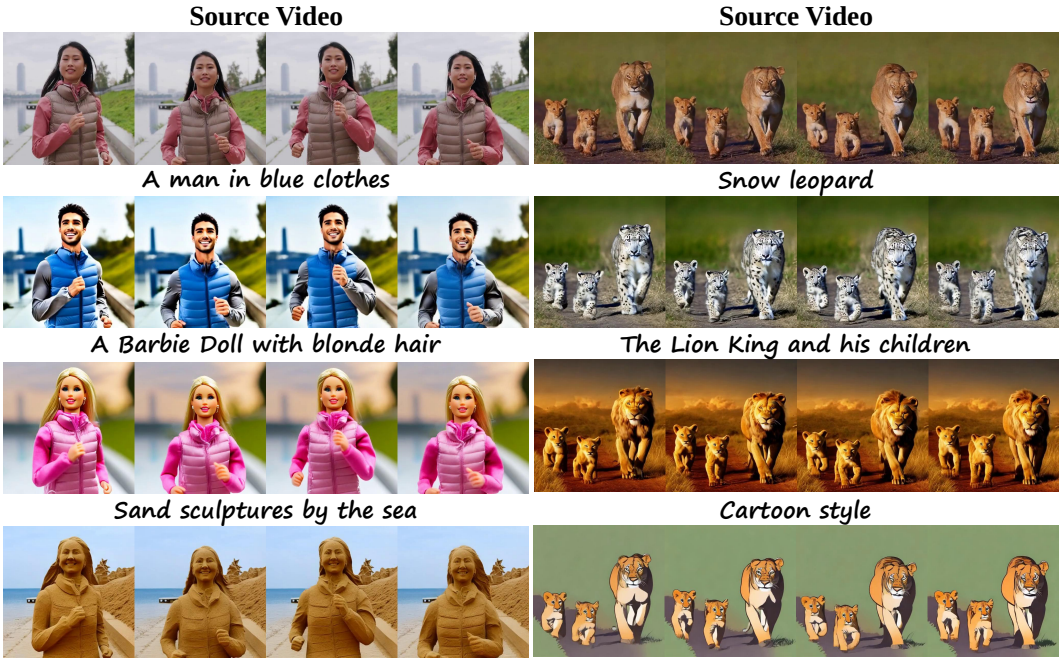

Figure 1: We propose **CO**rrespondence-guided **V**ideo **E**diting (COVE), which leverages the correspondence information of the diffusion feature to achieve consistent and high-quality video editing. Our method is capable of generating high-quality edited videos with various kinds of prompts (style,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

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3061

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 [27, 63, 65] have shown exceptional performance in image generation [57], thereby inspiring their application in the field of image editing [6, 25, 7, 53, 67, 24]. These approaches typically leverage a pre-trained Text-to-Image (T2I) stable diffusion model [57], using DDIM [64] 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 [57], 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 [26, 29, 62]. Consequently, there is a growing focus on adapting the pre-trained T2I diffusion for video editing [16, 33, 11, 72, 73, 54]. 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 [33, 54] or blurring [16] in generated videos), requiring additional components [30, 72, 11, 73] or needing extra training or optimization [73, 69, 41], etc.

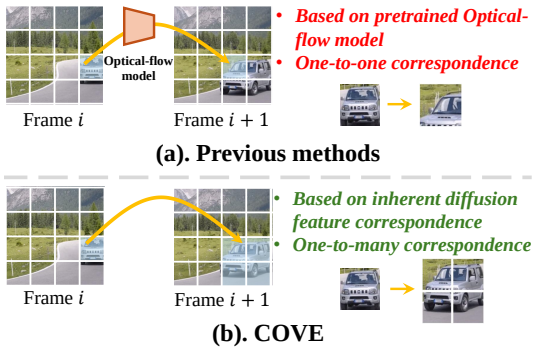

Figure 2: **Comparison between COVE (our method) and previous methods**[11, 73].

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 video are expected to share similar motion and semantic layout) determines that it naturally exists in the source video. Some previous methods [11, 73] 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 [32]. 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).

https://doi.org/10.52202/079017-3061

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 [66]. 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 [27, 63, 65] have recently showcased impressive results in image generation, which generates the image through gradual denoising from the standard Gaussian noise[12, 13, 52, 21, 57, 64, 22]. A large number of efforts on diffusion models [28, 34, 60] has enabled it to be applied to numerous scenarios [3, 15, 35, 39, 44, 49, 50, 14, 58, 9, 23, 47, 42, 18]. With the aid of large-scale pretraining [55, 61], text-to-image diffusion models exhibit remarkable progress in generating diverse and high-quality images [51, 71, 56, 57, 59, 19, 48, 18]. ControlNet [75] 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, 26, 29, 70, 10]. Despite successes in unconditional video generation [29, 74, 45], 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 [43, 68, 69, 46, 20], 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 [54] proposes to use attention blending to achieve high-quality edited videos while struggling to edit long videos. TokenFlow [16] reduces the effects of flickering through the linear combinations between diffusion features, while the smoothing strategy can cause blurring in the generated video. RAVE [33] 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 [73, 72, 11, 30]. Although the effectiveness of them, all of them severely rely on a pre-trained optical-flow model. Recent works [66] illustrate that the diffusion feature contains rich correspondence information. Although VideoSwap [17] 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

https://doi.org/10.52202/079017-3061

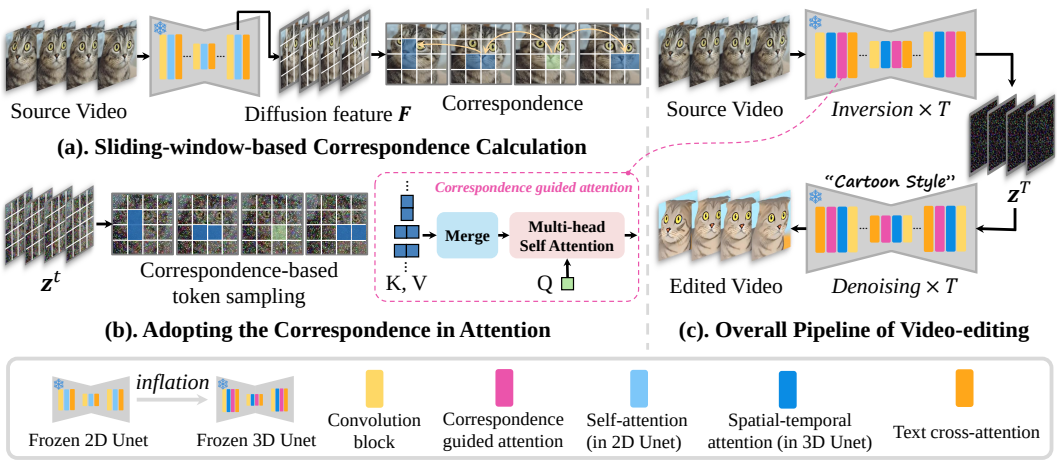

Figure 3: **The overview of COVE.** (a). Given a source video, we extract the diffusion feature of each frame using the pre-trained T2I model and calculate the correspondence among tokens (detailed in Figure 4). (b). During the video editing process, we sample the tokens in noisy latent based on correspondence 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 [27] 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 $\beta_t \in (0, 1)$.

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

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 [57] is proposed to address this issue by performing these processes in the latent space of a VAE [37].

DDIM Inversion. DDIM can convert random noise to a deterministic $\boldsymbol{x}_0$ during sampling [64, 13]. 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.

https://doi.org/10.52202/079017-3061

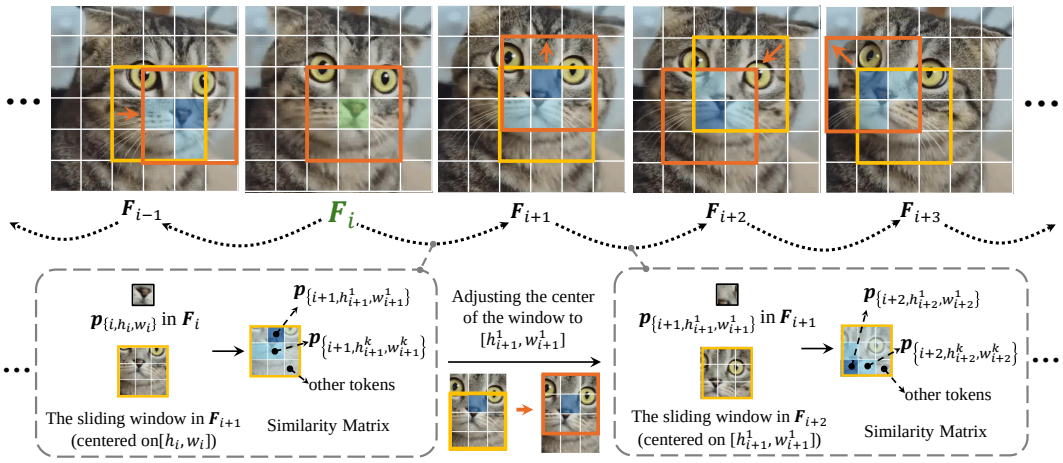

Figure 4: **Sliding-window-based strategy for correspondence calculation.** represents the token $\boldsymbol{p}_{\{i,h_i,w_i\}}$. and represents the obtained corresponded tokens in other frames.

Diffusion Feature Extraction. Given a source video $\boldsymbol{V}$ with N frames, a VAE [37] 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 $\boldsymbol{Z}^t = \{\boldsymbol{z}^t_1, \cdots, \boldsymbol{z}^t_N\}$ 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 [66]:

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

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.

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

Here the top-k-argmax$(\cdot)$ denotes the operation to find coordinates of the top K biggest values in a matrix, where $k \in \{1, \cdots, K\}$. $[h^k_j, w^k_j]$ 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

https://doi.org/10.52202/079017-3061

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 [73, 11], we also retain the spatial-temporal attention [69] 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 [11, 16]. 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.

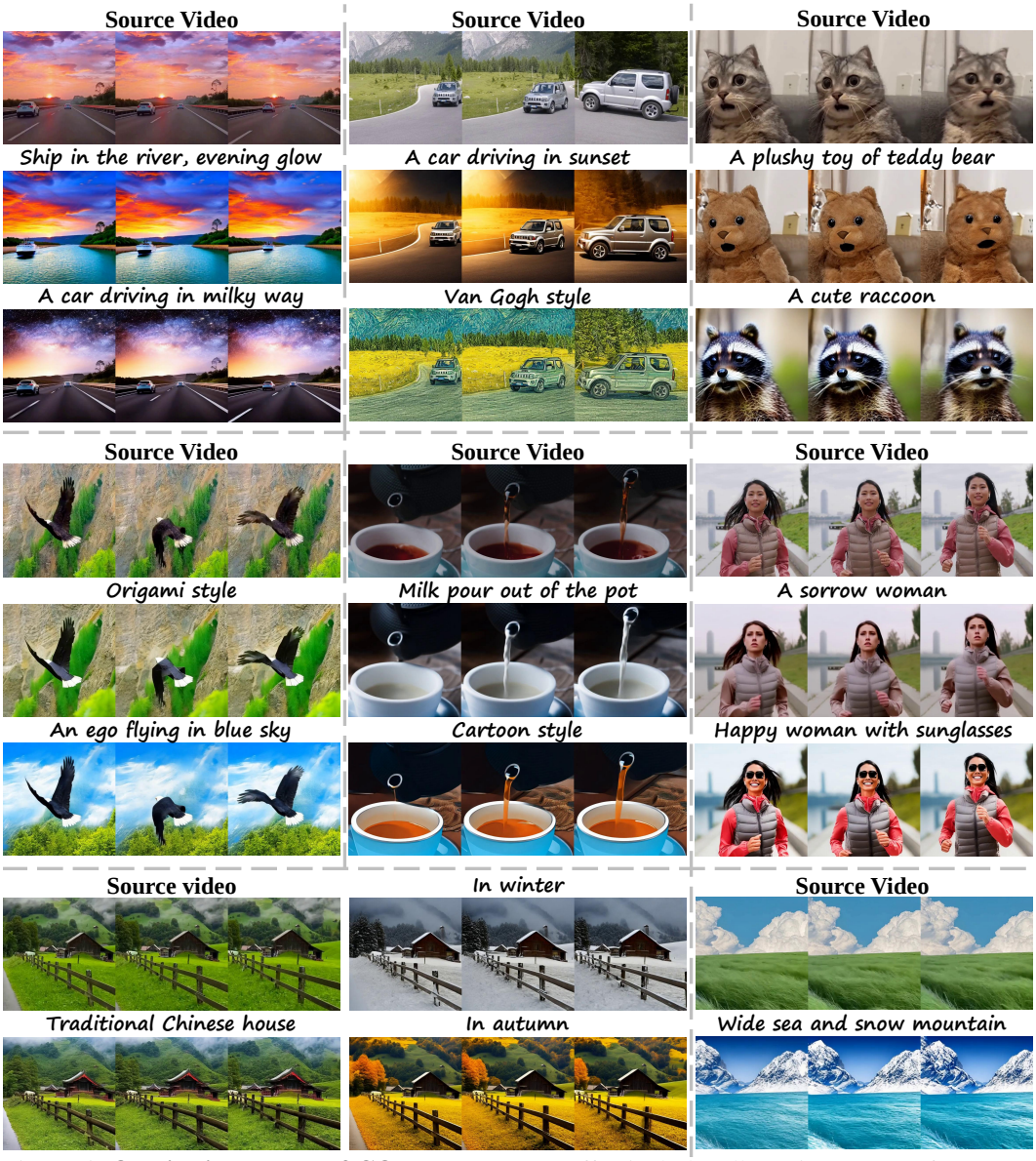

Figure 5: **Qualitative results of COVE.** COVE can effectively handle various types of prompts, generating high-quality videos. For both global editing (e.g., style transferring and background editing) and local editing (e.g., modifying the appearance of the subject), COVE demonstrates outstanding performance. 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 [66]. 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

https://doi.org/10.52202/079017-3061

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 [54], TokenFlow [16], FLATTEN [11], FRESCO [73] and RAVE [33]. 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 [31], 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

video by calculating the similarity of DINO [8] feature across frames. Motion Smoothness utilizes the motion priors of the video frame interpolation model [40] to evaluate the smoothness of the motion in the generated video. Aesthetic Quality uses the LAION aesthetic predictor [38] 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 [36] 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 [73]. 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 [54]	0.9622	0.9547	0.6258	0.6951	7.4%
TokenFlow [16]	0.9513	0.9803	0.6904	0.7354	13.0%
FLATTEN [11]	0.9617	0.9622	0.6544	0.7155	14.8%
FRESCO [73]	0.9358	0.9737	0.6582	0.6331	9.2%
RAVE [33]	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[31] 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.

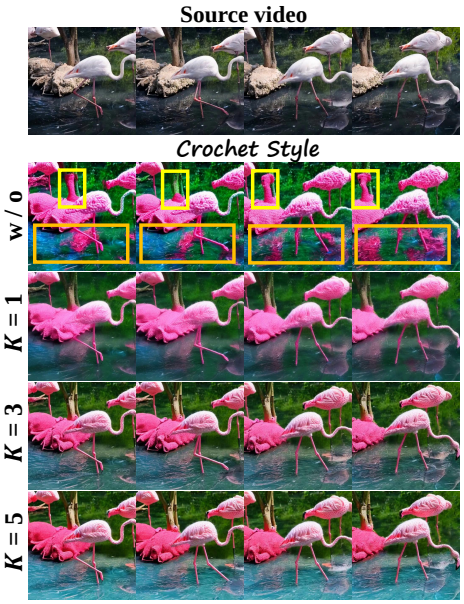

Figure 7: Ablation study about the correspondence-guided attention and the value of K. w / o means do not apply correspondence-guided attention.

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.

https://doi.org/10.52202/079017-3061

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 G.

Acknowledgments and Disclosure of Funding

This work was supported by the STI 2030-Major Projects under Grant 2021ZD0201404.

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, Q., Ma, Y., Wang, H., Yuan, J., Zhao, W., Tian, Q., Wang, H., Min, S., Chen, Q., Liu, W.: Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. arXiv preprint arXiv:2409.01055 (2024) 3

[10] 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

[11] 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

[12] 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

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

[14] Fang, C., He, C., Xiao, F., Zhang, Y., Tang, L., Zhang, Y., Li, K., Li, X.: Real-world image dehazing with coherence-based label generator and cooperative unfolding network. arXiv preprint arXiv:2406.07966 (2024) 3

[15] 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

[16] 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

[17] 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

[18] Guo, H., Dai, T., Ouyang, Z., Zhang, T., Zha, Y., Chen, B., Xia, S.t.: Refir: Grounding large restoration models with retrieval augmentation. arXiv preprint arXiv:2410.05601 (2024) 3

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

 https://doi.org/10.52202/079017-3061

[26] 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

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

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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

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

[36] 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

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

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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

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

[61] 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

 https://doi.org/10.52202/079017-3061

[62] 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

[63] 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

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

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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

[72] 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

[73] 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

[74] 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

[75] 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 [66], 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.

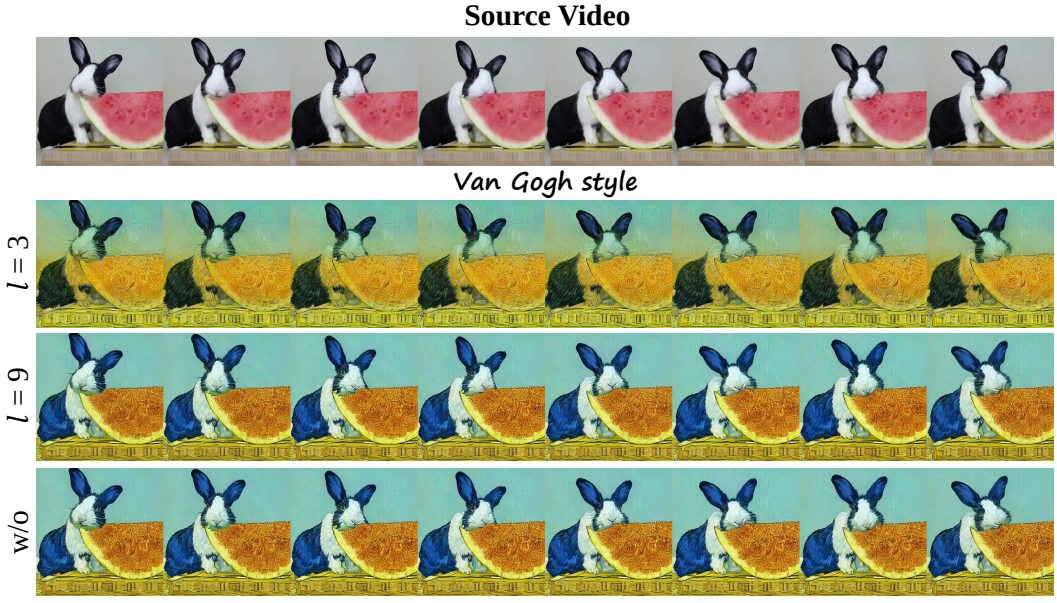

Figure 9: **Ablation Study on the window size** l**.**

https://doi.org/10.52202/079017-3061

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. The result illustrates that our method can effectively identify the corresponding tokens.

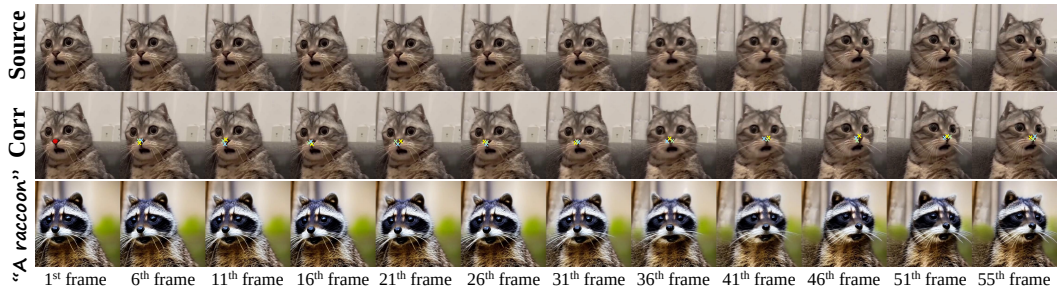

Figure 10: **Visualization of the correspondence in long videos.** Given a long video, we first obtain the correspondence information ($K = 3$) through the sliding-window strategy. Then, considering a point in the first frame (the red point in the first image of the second row), we visualize the correspondence (respectively marked in yellow, green, and blue) in each frame.

D Accuracy of Correspondance

The correspondence acquired through the diffusion feature is accurate and robust. As there is no existing video dataset with the annotated keypoints on each frame, to further evaluate its accuracy quantitatively, we collect 5 videos with 30 frames and 5 videos with 60 frames and manually label some keypoints on each frame. Then we report the percentage of correct keypoints (PCK).

Specifically, for each video, given the first frame with the keypoints, we obtain the predicted corresponding keypoints on other frames through the diffusion feature. Then we evaluate the distance between the predicted points and the ground truth. The predicted point is considered to be correct if it lies in a small neighborhood of the ground truth. Finally, the total number of correctly predicted points divided by the total number of predicted points is the value of PCK. The result in Appendix D illustrates that the diffusion feature can accurately find the correct position in most cases for video editing.

Method	PCK
Optical-flow Correspondence	0.87
Diffusion feature Correspondence	**0.92**

Table 5: **Accuracy of Correspondance.**

E Effectiveness of correspondence guided attention during inversion

The quality of noise obtained by inversion can significantly affect the final quality of editing. The Correspondence-Guided Attention (CGA) during inversion can increase the quality and temporal consistency of the obtained noise, which can further help to enhance the quality and consistency of edited videos. The ablation of it is shown in Fig. 11

F 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

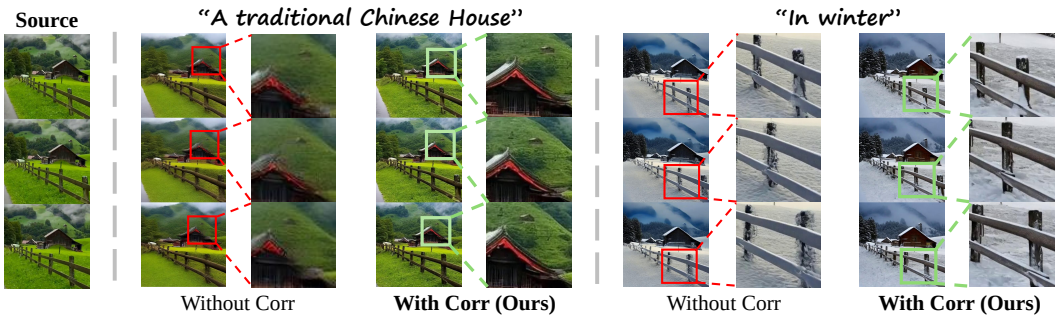

Figure 11: **Ablation Study about correspondence in inversion.** Here *Without Corr* means not applying the correspondence-guided attention during inversion, which suffers blurring and flickering. *With Corr* means the correspondence-guided attention is applied in both inversion and denoising stages, illustrating satisfying performance.

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.

G 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.

H More Qualitative Results

We provide more qualitative results of our method to illustrate its effectiveness, which is shown in Fig. 13 and Fig. 12.

https://doi.org/10.52202/079017-3061

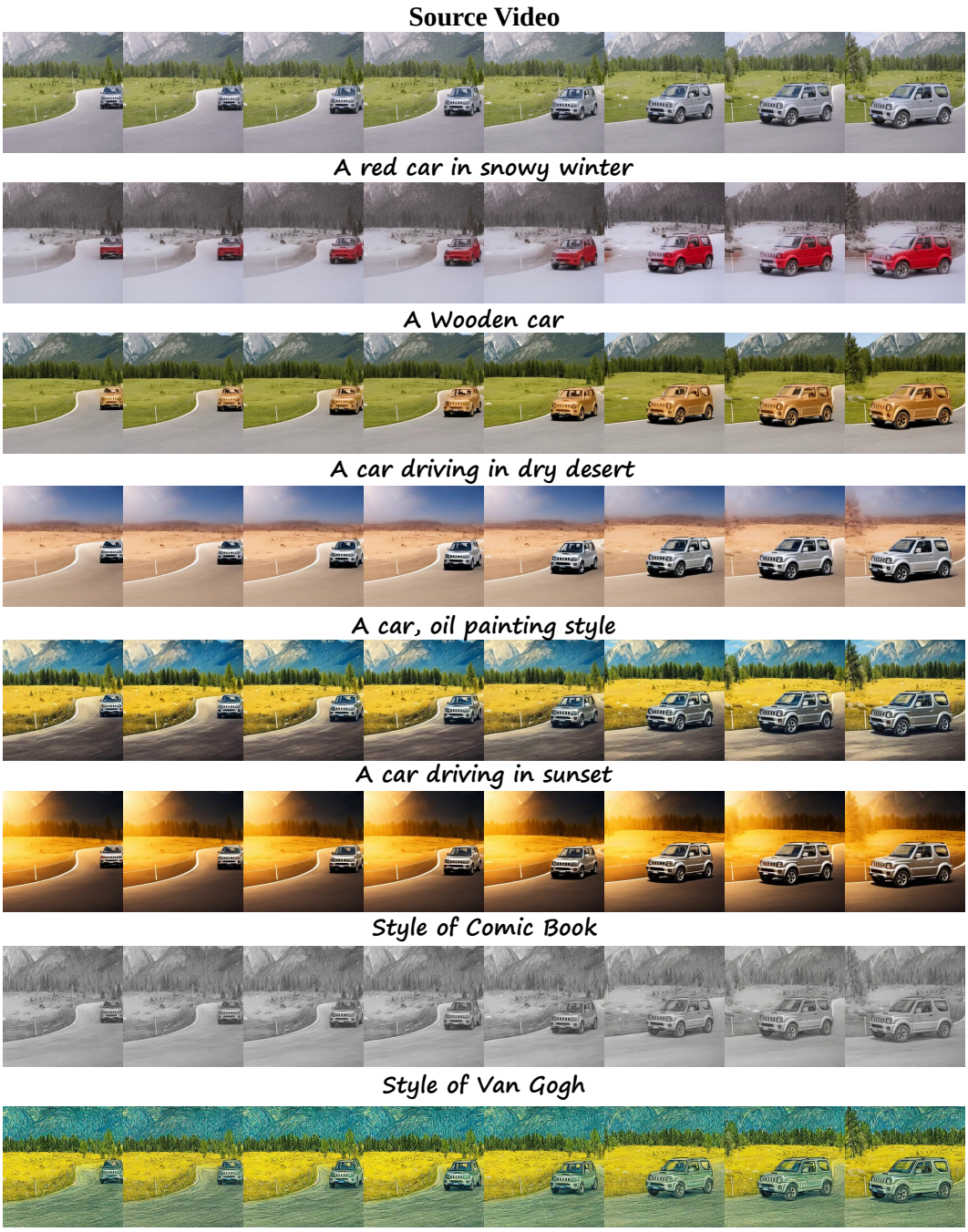

Figure 12: **Qualitative results of our methods.** Our method can effectively handle various kinds of prompts, generating high-quality videos. Results are best viewed in zoomed-in.

https://doi.org/10.52202/079017-3061

Figure 13: **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.

https://doi.org/10.52202/079017-3061

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of our works.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We prove that the self-attention can enhance the similarity among tokens, which is included in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the experimental details and submit the code in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3061

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit the code in the supplementary results while not creating a public GitHub repo.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiment details are specified.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The qualitative results are important in the field of generation. The error bar is relatively not necessary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have described the resources required to perform our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our work conform with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss this in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-3061

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and the license and terms of use explicitly are mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets introduced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No such experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-3061