Edit-A-Video: Single Video Editing with Object-Aware Consistency

Chaehun Shin* CHAEHUNY@SNU.AC.KR
Heeseung Kim* GMLTMD789@SNU.AC.KR
Che Hyun Lee SAGA1214@SNU.AC.KR
Sang-gil Lee TKDRLF9202@SNU.AC.KR
Data Science and AI Lab, ECE, Seoul National University, Seoul 08826, Korea
Sungroh Yoon† SRYOON@SNU.AC.KR
Data Science and AI Lab, ECE and Interdisciplinary Program in AI, Seoul National University, Seoul 08826, Korea

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

With advancements in text-to-image (TTI) models, text-to-video (TTV) models have recently been introduced. Motivated by approaches on TTV models adapting from diffusion-based TTI models, we suggest the text-guided video editing framework given only a pretrained TTI model and a single <text, video> pair, which we term **Edit-A-Video**. The framework consists of two stages: (1) inflating the 2D model into the 3D model by appending temporal modules and tuning on the source video (2) inverting the source video into the noise and editing with target text through attention map injection. Each stage enables the temporal modeling and preservation of semantic attributes of the source video. One of the key challenges for video editing is a background inconsistency problem, where the regions unrelated to the edit suffer from undesirable and inconsistent temporal alterations. To mitigate this issue, we also introduce a novel mask blending method, termed as temporal-consistent blending (TC Blending). We improve previous mask blending methods to reflect the temporal consistency, ensuring that the area where the editing is applied exhibits smooth transition while also achieving spatio-temporal consistency of the unedited regions. We present extensive experimental results over various types of text and videos, and demonstrate the superiority of the proposed method compared to baselines in terms of background consistency, text alignment, and video editing quality. Our samples are available on https://editavideo.github.io.

Keywords: Diffusion-based Generative Model, Text-based Video Editing

1. Introduction

Recently, generative models have made remarkable progress across various domains. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020), in particular, have shown state-of-the-art generation performance across multiple tasks, including text-to-image (TTI) generation (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022). In addition to generating images from text prompts, several works have used a pretrained TTI model for extended applications, such as personalized text-to-image (Ruiz et al., 2022; Gal et al., 2022;

∗. Equal Contribution, †. Corresponding Author.

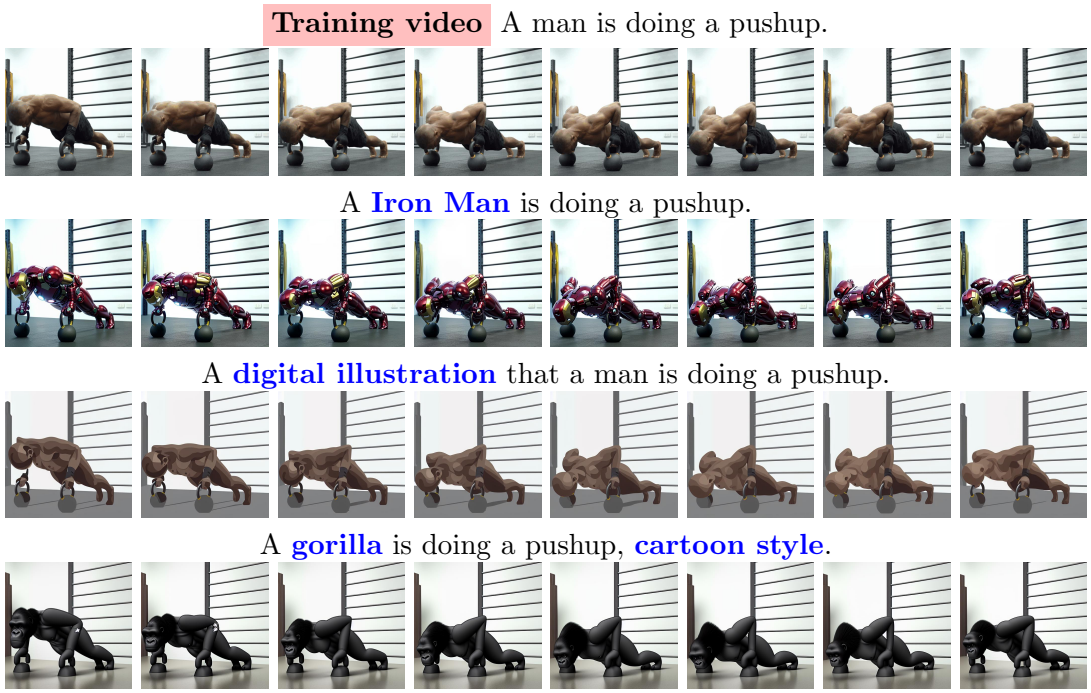

Figure 1: Edit-A-Video performs text-guided video editing from a single <text, video> pair and a text-to-image model.

Kumari et al., 2022) or text-guided image editing (Hertz et al., 2022; Mokady et al., 2022; Kawar et al., 2022).

Inspired by the success of the diffusion-based TTI models, various works have extended their output modality to videos, including text-to-video (TTV) generation and text-guided video editing. Among them, text-guided video editing models perform editing by leveraging the prior knowledge of the TTV model (Molad et al., 2023; Esser et al., 2023) or the inflated TTI model that captures temporal information (Wu et al., 2022). Several works (Wu et al., 2022; Ceylan et al., 2023) generate the whole video while editing, resulting in the editing of unwanted areas. Other methods (Liu et al., 2023; Qi et al., 2023) have proposed using image editing techniques to specify areas to be edited within each frame. However, these approaches have limitations in that they do not consider temporal information when calculating the editing regions for each frame.

In this paper, we propose Edit-A-Video, a framework designed for achieving temporal consistency in text-guided video editing. Edit-A-Video is a two-stage framework (see Fig. 2 for illustration). In the first stage, a pretrained 2D TTI model is inflated to a 3D TTV model with attention modules for spatio-temporal modeling, which is finetuned using a single source video. In the second stage, the source video is edited to match the target text description by inverting the source video to the Gaussian noise and injecting attention maps from the source to the target along the denoising process.

As previously mentioned, one of the main challenges in diffusion-based video editing is the accurate handling of the target object and the background. We observe that the edited

video suffers from a *background inconsistency problem*, where the edited video contains abrupt temporal changes in the background, which significantly degrades the quality. To tackle this issue, we propose a novel blending method tailored to the task, called *temporal-consistent blending (TC Blending)*. TC Blending extends a spatial local blending technique originally proposed for a still image (Hertz et al., 2022) and automatically generates a sharp, spatio-temporally consistent blending mask that closely approximates the region to be edited in the source video.

Together with the proposed TC Blending method, Edit-A-Video can effectively generate a realistic video that matches the target text description and captures the dynamic actions of the source video ensuring a smooth transition, while also maintaining the spatio-temporal consistency of the background. We carry out extensive experiments on one-shot editing over various videos and prompts, and demonstrate that Edit-A-Video is the most favorable method compared to baselines through subjective human preference study along with qualitative analysis. We further conduct an in-depth analysis by comparing to baselines based on automatic evaluation of numerical metrics, which shows the improvement of consistency and text alignment. We also demonstrate the effectiveness of the proposed TC Blending in terms of consistency and overall editing quality through ablation studies.

In summary, our key contributions are as follows:

- This study presents Edit-A-Video, a one-shot video editing framework that effectively combines a pretrained text-to-image model and editing techniques suitable for video editing.
- Along with the extension of image editing techniques to video, we propose a novel blending method called temporal-consistent blending (TC Blending), alleviating the background inconsistency problem.
- We analyze the effect of injecting each attention map for spatio-temporal consistency in our framework.

2. Background

2.1. Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020) is a type of probabilistic generative model that consists of a forward process that gradually transforms data into $z \sim N(0, I)$ and a reverse process, the opposite trajectory. Following notation in Ho et al. (2020), the forward process of diffusion is defined as follows:

$$\begin{aligned} q(x_t|x_{t-1}) &:= N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \\ x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \end{aligned} \tag{1}$$

where $z_t \sim N(0, I)$, β_t is a user-defined noise schedule, and $\bar{\alpha}_t := \prod_{i=1}^{t}(1-\beta_i)$. Using Eq. 1, we can obtain the posterior $q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t)$.

For sampling data from noise, we define a reverse process that gradually transforms noise into data for sampling. Since $q(x_{t-1}|x_t)$ is intractable, Ho et al. (2020) approximate

this term to the transition network $p_\theta(x_{t-1}|x_t)$, which is defined in a Gaussian form as below.

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$
$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \tag{2}$$

where $\Sigma_\theta(x_t, t)$ can be a fixed value (Ho et al., 2020) or trainable parameters (Nichol and Dhariwal, 2021), and $\epsilon_\theta(x_t, t)$ can be trained by minimizing $\mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$ (Ho et al., 2020). On the contrary, DDIM (Song et al., 2020) proposes to set $\Sigma_\theta(x_t, t)$ to 0 for a subset of the reverse diffusion process, which enables a deterministic and fast sampling. In this paper, we follow the formulation of DDIM (Song et al., 2020).

2.2. Tune-A-Video

Compared to text-to-image generation, text-to-video generation (Singer et al., 2022; Ho et al., 2022; Esser et al., 2023) is a data-hungry task since the training of TTV model requires a large-scale <text, video> paired dataset. Recently, Tune-A-Video (Wu et al., 2022) successfully trained a TTV model with a single video leveraging prior knowledge of a pretrained TTI model, while maintaining consistency between frames. To generate a temporally-coherent video, Tune-A-Video inflated 3x3 2D convolution layers from TTI models to 1x3x3 3D convolution layers and appended additional temporal attention (T-Attn) modules. To enhance the temporal consistency more, Tune-A-Video replaces spatial self-attention with a novel attention layer named as sparse spatio-temporal attention (ST-Attn). Instead of full frame attention which computes the attention for every other frame, ST-Attn computes the attention of current frames only on the first and the previous frame. It reduces the computational cost of full frame attention which is quadratic with respect to the number of frames to linear cost, while maintaining the temporal consistency in video generation. By tuning three types of attention layers (T-Attn, ST-Attn, and cross-attention between text and video) from the inflated model, Tune-A-Video is capable of generating temporally-coherent videos.

Despite its ability to synthesize video with temporal consistency, Tune-A-Video fails to maintain the contents that should remain unedited since it generates entire video frames using a modified text prompt. For example, the background regions of the video generated by Tune-A-Video are modified and are not preserved with respect to the source video. In contrast, Edit-A-Video edits a video with given text prompts while maintaining the attributes unrelated to the modified text based on the novel blending method.

2.3. Real Image Editing with Null-Text Inversion

Text-guided image editing is a task to edit the content of a given image only specified by the text prompts. While TTI models can produce high-quality images from the given text prompts, the model tends to generate completely different images as the source text is modified to serve as the editing target. Hertz et al. (2022) observe that the cross-attention maps between the image feature and the source text reflect the spatial layout of the source image. Based on the remarkable property of cross-attention maps, Hertz et al. (2022) propose a novel framework, Prompt-to-Prompt (PTP), for image editing which injects the attention maps of the source image into the generation process conditioned on the target text. PTP

further enhances the degree of preservation of the source image by approximating the desired editing target region from the cross-attention maps and blending the target region of the edited image with the source image.

Although PTP is a strong framework for synthetic image editing, a latent vector corresponding to the real image is required for real image editing. DDIM inversion is one of the methods to invert the real image to a latent vector. However, DDIM trajectory is severely distorted when a classifier-free guidance (Ho and Salimans, 2021) is applied through the reverse process, which results in the poor reconstruction quality of the real image. As an effective approach for the limitation, Null-Text Inversion (NTI) (Mokady et al., 2022) optimizes the null-text embedding which is used in classifier-free guidance so that the reverse process trajectory with classifier-free guidance does not deviate from the DDIM inversion trajectory. NTI results in the near-perfect reconstruction of the image even under classifier-free guidance. Together with the PTP framework, NTI shows that text-guided image editing can also be applied to real image.

Despite the success of text-guided image editing, image editing methods cannot be directly applied to video editing. Compared to image editing, maintaining temporal consistency is crucial when it comes to video editing. We hence append sparse spatio-temporal attention (ST-Attn) and temporal attention (T-Attn) on top of NTI and PTP framework and propose a novel blending method called temporal-consistent blending (TC Blending) for successful video editing.

2.4. Text-Guided Video Editing

Recently, text-based image creations expand their domain from the image to the video. Text-to-Video models synthesize high-quality videos corresponding to a given text. Likewise, following the text-guided image editing methods, various approaches for text-guided video editing have also been proposed. Text-guided video editing modifies the given source video to reflect the target text while preserving some content from the source video. Earlier works (Esser et al., 2023; Molad et al., 2023) edit the video by exploiting the TTV model trained on a large-scale <text, video> paired dataset, which is computationally challenging to most practitioners. Other approaches (Bar-Tal et al., 2022; Lee et al., 2023) utilize the neural layered atlas (NLA) of the source video for editing. While they enable video editing by applying image editing methods to the NLA of the video, acquiring NLA is time-consuming and lacks efficiency.

In contrast to prior approaches, there are several works that leverage the TTI model to perform text-guided video editing similar to ours. Some previous works (Wu et al., 2022; Ceylan et al., 2023; Wang et al., 2023) generate the entire video frames including the area not related to the target object during editing, which makes it difficult to maintain the content of the source video. Similar to Edit-A-Video, some concurrent works (Qi et al., 2023; Liu et al., 2023) perform editing through attention map control. These methods improve editing performance by blending the source video and target region specified by the target text using a mask extracted from the cross-attention map. However, the cross-attention map is calculated between text and each frame and does not consider temporal information between frames. Thus, the frame-wise blending mask extracted from cross-attention map can be temporally inconsistent, resulting in undesirable artifacts including abrupt changes

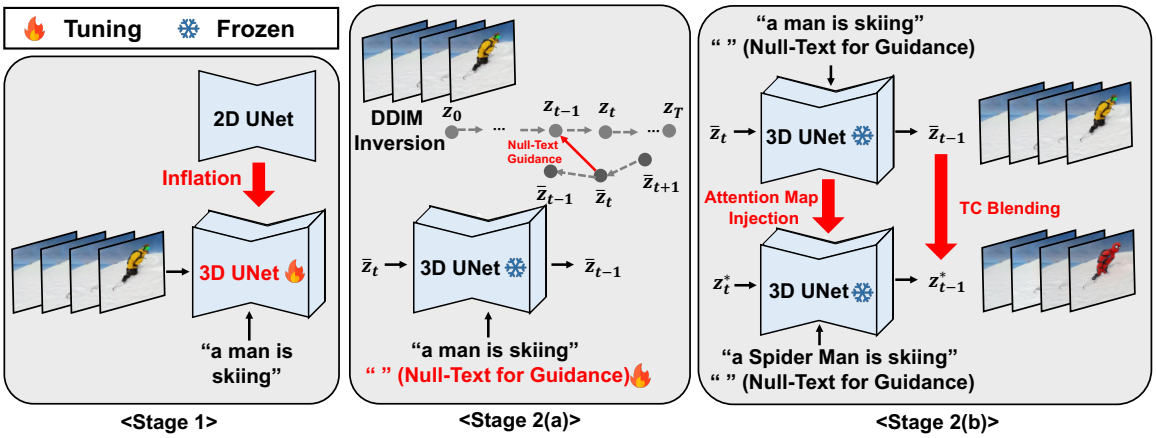

Figure 2: **Overall Editing Procedure of Edit-A-Video** In stage 1, Edit-A-Video inflates the 2D UNet into 3D model by appending the temporal module and evolving the 2D conv into 3D conv and self-attention into sparse spatio-temporal attention. Then, the source video is inverted to a specific Gaussian noise by DDIM inversion, and null-text embedding is optimized so that the source video can be reconstructed along with the null-text guidance in stage 2(a). Finally, in stage 2(b), the edited video is generated from the inverted noise and target text through attention map injection, TC Blending, and optimized null-text embedding.

in the background. To mitigate the aforementioned issue, Edit-A-Video proposes TC blending, a blending method that is capable of extracting sharp and frame-consistent masks by considering temporal information from sparse spatio-temporal attention.

3. Method

In this section, we introduce Edit-A-Video, a framework designed to consistently edit a given video into a desired object or style using a diffusion-based TTI model. In Sec. 3.1, we formulate our 2-stage editing procedure, extending TTI model to learn the temporal relationship and editing the contents of the source video according to the target prompt textcolorredvia attention map injection. To achieve more consistent editing frame by frame, we propose a temporal-consistent blending (TC Blending) method in Sec. 3.2. Finally, we discuss the role and effect of three types of attention modules in our methodology in Sec. 3.3.

3.1. Framework

As shown in Fig. 2, Edit-A-Video follows a two-step process to edit the given video corresponding to a target prompt. In the first stage, we inflate the TTI model to TTV model using the method of Tune-A-Video (Wu et al., 2022). Unlike the TTI model, which has two types of attention (self-attention in images and cross-attention between text and image), our inflated TTV model has three types of attention: cross-attention, temporal attention, and

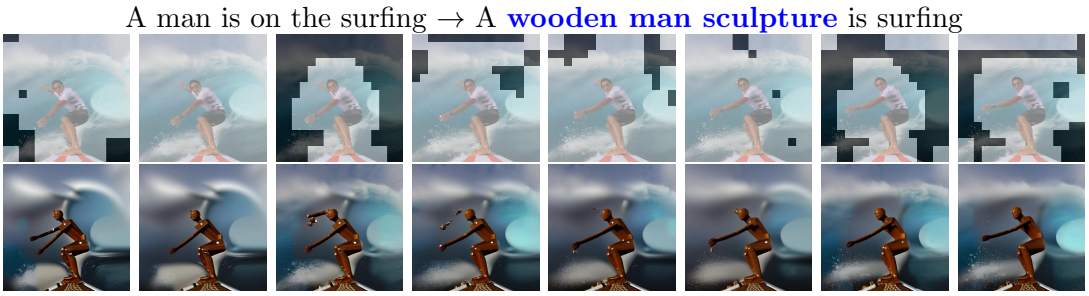

Figure 3: **Background Inconsistency Problem** The spatial local blending mask originally proposed in Hertz et al. (2022) cannot consider the temporal consistency, resulting in temporally variant background after editing, which can be identified in the waves close to the wooden sculpture.

sparse spatio-temporal attention. We only train these attention modules in the inflated 3D TTV model using a single video, ensuring frame-by-frame consistency during video editing.

In the second stage, we extract the inversion trajectory $\{z_t\}_{t=1}^{T}$ starting from the source video to the Gaussian noise z_T and train the null-text embeddings so that the generation trajectory from z_T is still close to the inversion trajectory under the classifier-free guidance. Starting from the latent variable z_T of the source video, we edit the video by injecting three types of attention maps of the source video into the generation process of edited video, extending the editing methods in the image domain (Hertz et al., 2022; Mokady et al., 2022). Since the inflated model has newly added attentions previously absent in the TTI model, we analyze the role and effect of each attention module and describe in Sec. 3.3.

However, when we edit the video as described above, we observe that unwanted region is edited inconsistently along the frames, which results in undesirable and abrupt artifacts. We call this issue the *background inconsistency problem*. To mitigate this issue, we analyze the cause and propose a *temporal-consistent blending*, which we call *TC Blending* for short, in the following section.

3.2. Temporal-Consistent Blending

When Edit-A-Video edits an object, only the specific region that correlates to the target object should be modified, while the rest of the video should be preserved. In previous work of image editing, PTP (Hertz et al., 2022) proposed the local blending method, which approximates a mask of the object region from the cross-attention map and performs editing only in the masked region. However, since the cross-attention of the inflated 3D model is computed frame-by-frame, the local blending mask from the cross-attention map considers only the spatial dimension and lacks modeling of temporal dependency. This indicates that the smoothness of blending masks across frames cannot be ensured, resulting in the potential occurrence of color-variant artifacts when a specific region is included in the blending mask of one frame but not in another. This background inconsistency problem is demonstrated clearly in Fig. 3, where the background region close to the target object of editing is severely distorted, and the frames are highly inconsistent across the temporal axis.

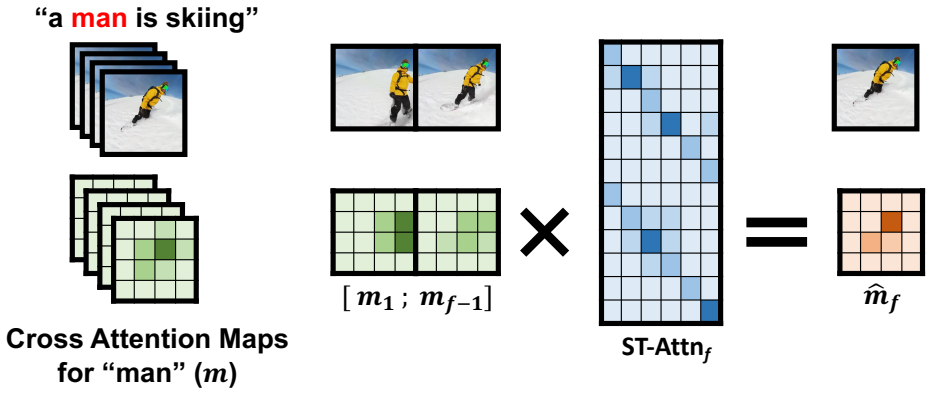

Figure 4: **TC Blending Mask Computation** The first and previous frames interact with the corresponding sparse spatio-temporal attention map and yield the new temporal-consistent blending mask.

To address this problem, we propose a novel blending method called temporal-consistent blending (TC Blending), which acquires a spatio-temporally consistent blending mask. For efficient temporal modeling, we use the sparse ST attention map in our 3D inflated model as a proxy for the interaction between the current frame mask and the first and previous frame masks. The sparse ST attention map is calculated by taking the current frame feature z_f as query, the first frame feature z_1 and the previous frame feature z_{f-1} as key:

$$\text{ST-Attn}_f = \text{Softmax}(QK^T/\sqrt{d}), Q = W^Q z_f, K = W^K[z_1; z_{f-1}], \tag{3}$$

where f is the index of current frame, [;] denotes the concatenation, and d is the feature dimension of the projection layer.

Intuitively, the sparse ST attention map of the current frame with the first frame ensures that the target object is maintained in every mask, while that with the previous frame enforces a smooth transition in the blending mask sequence. With TC Blending, it is possible to achieve a sharp blending mask that accurately detects the target object in the frame and ensures a smooth transition in the blending masks, reducing background inconsistencies.

Following Hertz et al. (2022), we first acquire the cross-attention maps $m \in \mathbb{R}^{F\times H\times W}$ according to the original word and new word as initial blending maps, where the F, H, W are the temporal and spatial dimension of the feature. Then, for each attention map m, we normalize it frame-wise to balance the scale between each frame and flattening:

$$\tilde{m} = \text{Flatten}(m/\sum_{H,W}(m)), \text{where } \tilde{m} \in \mathbb{R}^{F\text{x}HW}. \tag{4}$$

Afterward, we encourage the interaction between the frame-wise maps by computing the weighted average of the first and previous map $[\tilde{m}_1; \tilde{m}_{f-1}] \in \mathbb{R}^{(HW\times 2)}$ with sparse ST attention map $\text{ST-Attn}_f \in \mathbb{R}^{(HW\times 2)\times(HW)}$ of the current frame, which is shown in Fig. 4. Then, we binarize it by thresholding:

$$\alpha_f = B(\hat{m}_f, \tau), \hat{m}_f = [\tilde{m}_1; \tilde{m}_{f-1}] \times \text{ST-Attn}_f, \tag{5}$$

where B is the binarizing function with threshold τ.

Similar to PTP, we utilize the union of binary blending masks for the original word and new word, denoted as α, as the final blending mask. This allows us to edit both the areas of the original object and the target object. We use this binary blending mask α to perform local editing, where we preserve the background outside the mask and only generate the contents inside the mask:

$$\hat{z}_t = \bar{z}_t \odot (1 - \alpha) + z_t^* \odot \alpha, \tag{6}$$

where $\bar{z}_t$ is the reconstruction of the source video, z_t^* is the edited video, and $\odot$ is element-wise multiplication.

3.3. Hyperparameters for Editing

The core idea of Prompt-to-Prompt (PTP) to preserve the spatial layout of input is by injecting 2D cross-attention maps and self-attention maps from pretrained TTI models. The duration of the injection process, a portion of timestep over the entire sampling step until which the attention map is injected, is a key factor in controlling the reflection ratio of the target prompt. Edit-A-Video, on the other hand, uses sparse spatio-temporal attention and temporal attention instead of self-attention. As a result, the effect of the duration of attention injection may differ from that of PTP. In this section, we describe the effect of injection for each type of attention, and the corresponding samples are visualized in the Supplementary Materials.

Cross-Attention The cross-attention layer performs an attention operation between text tokens and frames, taking into account the spatial layout of each text token in the frames. There is a trade-off depending on the duration of the injection phase. If injection occurs only at the beginning of generation, the generated frames are strongly conditioned on the target text prompt, but hard to maintain the spatial layout. On the other hand, if injection occurs throughout the entire generation process, the spatial layout of the source video is well preserved, yet it does not include the concepts from the text prompt. Empirically, we found that the duration of 0.2 is sufficient to retain the spatial layout of the source video, while the generated video represents the semantics of the target text.

Temporal Attention The temporal attention (T-Attn) layer is an additional attention module computed along the time axis of a video to model the temporal relationship between frames, yet it does not model spatial dependency. We found that T-Attn maps are distributed uniformly and that the duration of injection does not have a significant impact. We set a duration value of 0.8 for temporal attention. However, varying this value does not have a notable impact on the editing quality, and the corresponding qualitative results are in Supplementary Materials.

Sparse Spatio-Temporal Attention The sparse spatio-temporal attention (ST-Attn) layer is an attention method designed for video, where the attention matrix of the current frame is calculated only on the first and previous frames. ST-Attn replaces the self-attention layer from the pretrained TTI models, where the model only requires the dependencies between pixels in a single image. In addition to temporal attention, ST-Attn improves temporal consistency by attending to other frames while maintaining efficiency by only visiting two frames. We observed that insufficient duration of ST-Attn results in the inability to adequately represent the dynamic action of the source video. On the contrary, an excessively

Table 1: **Qualitative Comparisons to Baselines** We measure the overall human preference score (User Score (O)) and automatic metric scores for the comparisons to baselines.

Method	User Score (O) (↑)	Text Alignment (↑)	LPIPS (↓)	PSNR (↑)
Edit-A-Video (Ours)	3.80 ± 0.10	30.2688	0.2625	20.0992
Tune-A-Video	3.46 ± 0.10	30.0514	0.4482	14.5753
SDEdit	3.40 ± 0.10	28.4203	0.2711	20.4767
Video-P2P	3.66 ± 0.10	30.0842	0.3047	17.5760

long duration tends to capture not only the actions of the source video but also the objects within it, which hinders the editing process toward the target object. Therefore, to maintain the dynamic of the source video and facilitate editing toward the target object, we set the duration of sparse spatio-temporal attention to 0.5.

4. Experiments

4.1. Implementation Details

We implement our method based on the stable-diffusion-v1-4[1], publicly available TTI model (Rombach et al., 2022). We finetune only the latent diffusion model in TTI model on 8 frame 512×512 video for 300 steps in temporal modeling and 500 steps in inversion, while fixing the autoencoder to encode each frame independently. At inversion and sampling, we use 50 step DDIM sampler and set the classifier-free guidance scale to 7.5. From the analysis introduced in Sec. 3.3, we use cross-attention injection duration as 0.2, spatio-temporal attention injection duration as 0.5, and temporal attention injection duration as 0.8. We set TC blending threshold (τ) as 0.25 and show the effects of the value in Sec. 4.3.

For the quantitative evaluation, we edit a total 100 <text, video> pairs (four captions for each of 25 videos), and the videos are collected from the web and DAVIS dataset (Pont-Tuset et al., 2017) as other works (Esser et al., 2023; Bar-Tal et al., 2022). We set two of the four sentences for each video to change the style including the background, and the other two sentences to change the object. We compare our model to baselines by human preference study, which we refer to as User Score (O), and automatic evaluation metrics. In the human preference study, we ask 62 users to grade the overall quality score of the edited video on a scale of $1-5$ considering three aspects: background preservation, text alignment, and video realism. We include detailed explanations in the Supplementary Materials.

For the detailed analysis, we evaluate all models with three automatic metrics, one for editing performance and two for background preservation. We measure the text alignment for the editing performance, which estimates how much the editing reflects the target text by averaging the cosine similarity between CLIP embedding of target text and CLIP image embeddings of all frames. We further measure the distance between the source video and the target video for the background preservation by LPIPS (Zhang et al., 2018) and PSNR following the previous works (Esser et al., 2023; Mokady et al., 2022; Hertz et al., 2022).

1. Stable Diffusion: https://github.com/CompVis/stable-diffusion

Figure 5: **Qualitative Results** Edit-A-Video outperforms in editing compared to other baselines.

Table 2: **TC Blending Ablation Study** We demonstrate TC Blending's impact through subjective scores (User Scores) and automatic metrics. User Score (O) represents overall editing quality, while User Score (P) assesses non-target content preservation, including the background. Mask IoU measures IoU between the foreground mask from the saliency detector and the blending mask.

Method	User Score (O) (↑)	User Score (P) (↑)	LPIPS (↓)	PSNR (↑)	Mask IoU (↑)
Edit-A-Video	3.80 ± 0.10	4.13 ± 0.13	0.2625	20.0992	0.3805
w/o TC-Bld	3.69 ± 0.10	3.96 ± 0.13	0.2723	19.8628	0.2371

4.2. Baseline Comparisons

We compare our method with three baselines quantitatively and qualitatively: (1) *Tune-A-Video*: generating the video from target prompt after tuning the inflated 3D model with source text-video pair. (2) *SDEdit*: Based on Tune-A-Video, editing the video with another method, SDEdit (Meng et al., 2021) which injects the noise to video until intermediate timestep $t_0 = 25$ among 50 steps following Meng et al. (2021) and denoises from it conditioned on target prompt. (3) *Video-P2P*: concurrent single video editing method, which inflates the 2D model by replacing self-attention with first-frame attention, where the attention matrix of the current frame is calculated only based on the first frame.

Quantitative Results We perform the user evaluation to let the participants grade the scores on the edited videos. Owing to the proposed techniques, Edit-A-Video achieves superior performance compared to baselines with statistical significance (p-value < 0.05 from the Wilcoxon signed-rank test), as shown in table 1. We further measure automatic evaluation metrics to support the user score, which are also included in table 1. Since Tune-A-Video generates entire frames of video corresponding to the target text, the synthesized sample accurately reflects the target text. However, we observe that Tune-A-Video modifies even the background to be preserved, which is in line with the results obtained from LPIPS and PSNR measurements. SDEdit, another baseline, preserves the contents of the source video, yet shows the lowest text alignment score, which indicates that it does not reflect the target text faithfully. Unlike the aforementioned two baselines, Video-P2P preserves the property that is independent of the target editing to some extent while reflecting the target text. Edit-A-Video exhibited superior performance compared to Video-P2P in all metrics. Through these results, we confirm that Edit-A-Video is capable of editing video correspond to target text while preserving details that should be remained.

Qualitative Results Fig. 5 presents qualitative results of ours and baselines. Tune-A-Video fails to maintain the background content, such as the color of the brick on the floor, while SDEdit fails to generate samples that reflect the target text. Although Video-P2P generates higher-quality samples than previous baselines, its imprecise blending mask leads to unintended changes in regions that should not be edited, such as the positioning of the legs. Compared to these baselines, Edit-A-Video generates samples that preserve the property that should remain unedited and align accurately with the target prompt. More examples are included in the Supplementary Materials.

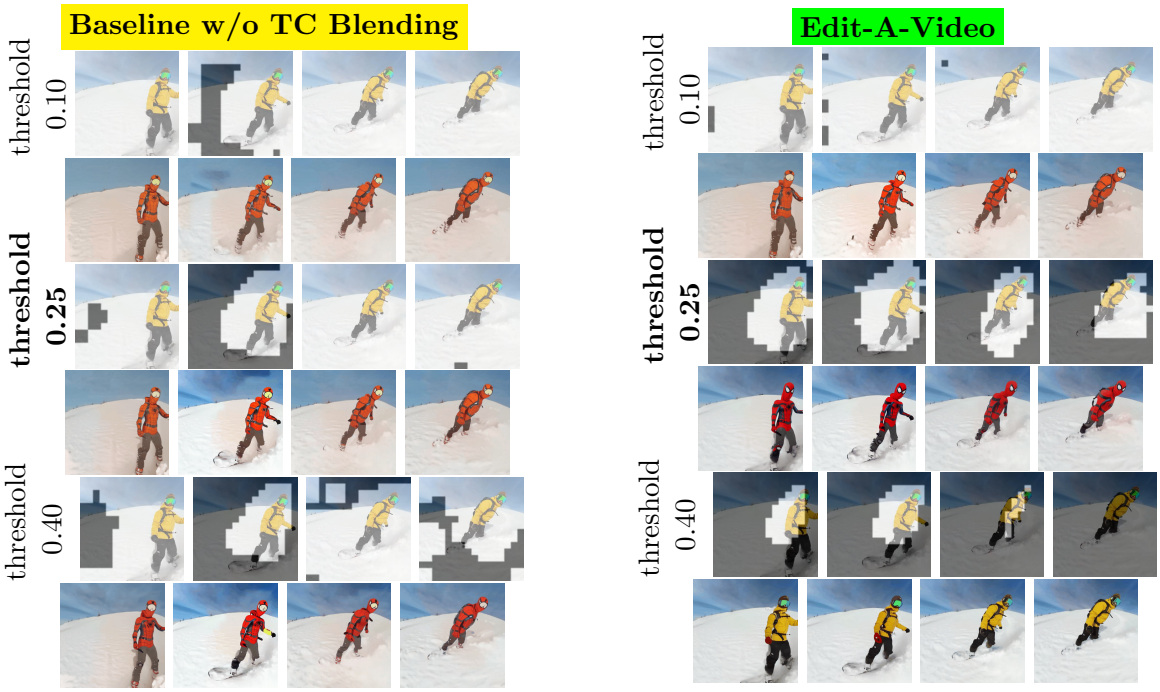

Figure 6: **Qualitative Ablation Studies** We visualize the editing results for the target text "A Spider Man is skiing". The samples on the left are based on the masking threshold without the proposed blending method applied, and the samples on the right are the results with the proposed blending method applied.

4.3. Ablation

In this section, we analyze the TC Blending mask's impact through an ablation study. We evaluate the editing targets of 50 out of the 100 pairs of <text, video> used in Sec 4.2, focusing specifically on object editing. We evaluate the proposed blending method's effectiveness using two human preference scores and three automatic metrics. The human preference scores, User Score (O) and User Score (P), are rated on a 5-point scale, measuring overall editing quality and non-target region preservation. For automatic metrics, we use LPIPS, PSNR, and a newly proposed metric, Mask Intersection over Union (Mask IoU). Mask IoU quantifies the overlap between the blending mask and the target object region, calculated as the Intersection over Union (IoU) between these two regions. To capture the target object region, we use a publicly available saliency detector (Zheng Peng and Huan, 2023).

As shown in table 2, Edit-A-Video achieves higher user scores than the model without TC Blending (p-value < 0.01 from the Wilcoxon signed-rank test). Furthermore, through various automatic metrics and qualitative results in Fig 6, we confirm that TC Blending is effective in accurate target object masking and background preservation.

We also demonstrate the effect of mask threshold value τ. In Fig 6, τ controls the editing region size. Without TC Blending, adjusting the threshold fails to capture sharp object regions effectively, causing abrupt frame-wise mask changes and undesirable artifacts. In

contrast, Edit-A-Video achieves sharp and smooth masks by properly setting the threshold. These results confirm that our proposed TC Blending mitigates background inconsistency.

5. Conclusion

We propose Edit-A-Video, the video editing framework only given the single <text, video> pair and pretrained text-to-image (TTI) model. Edit-A-Video inflates the TTI model and tunes the model on a given source video for the temporal modeling, and enables editing the video with target prompt by the inversion and attention map injection. We also suggest the temporal-consistent blending method which ensures content preservation and temporal coherence by making use of the temporal modeling capability in the model. Our framework achieves superior performance to the baselines in various aspects. We anticipate that this method will provide a simple and intuitive video editing method.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation grant funded by the Korea government (MSIT) [2021-0-01343, AI Graduate School Program (SNU)], National Research Foundation of Korea grant funded by MSIT (2022R1A3B1077720), and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, SNU in 2023.

References

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022.

Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

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

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

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14317–14326, June 2023.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

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

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

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Learning Representations*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, October 2020. URL https://arxiv.org/abs/2010.02502.

Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Wang Shuo Xiang Tian-Zhu Zheng Peng, Qin Jie and Xiong Huan. Memory-aided contrastive consensus learning for co-salient object detection. In *AAAI*, 2023.