Re-Attentional Controllable Video Diffusion Editing

Yuanzhi Wang[1,2], **Yong Li**[1,3], **Mengyi Liu**[2], **Xiaoya Zhang**[1,*],
Xin Liu[4], **Zhen Cui**[1,*], **Antoni B. Chan**[3]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2]Department of Content Security, Kuaishou Technology, Beijing, China
[3]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[4]SeetaCloud, Nanjing, China

Abstract

Editing videos with textual guidance has garnered popularity due to its streamlined process which mandates users to solely edit the text prompt corresponding to the source video. Recent studies have explored and exploited large-scale text-to-image diffusion models for text-guided video editing, resulting in remarkable video editing capabilities. However, they may still suffer from some limitations such as mislocated objects, incorrect number of objects. Therefore, the controllability of video editing remains a formidable challenge. In this paper, we aim to challenge the above limitations by proposing a *Re-Attentional Controllable Video Diffusion Editing* (***ReAtCo***) method. Specially, to align the spatial placement of the target objects with the edited text prompt in a training-free manner, we propose a Re-Attentional Diffusion (RAD) to refocus the cross-attention activation responses between the edited text prompt and the target video during the denoising stage, resulting in a spatially location-aligned and semantically high-fidelity manipulated video. In particular, to faithfully preserve the invariant region content with less border artifacts, we propose an Invariant Region-guided Joint Sampling (IRJS) strategy to mitigate the intrinsic sampling errors w.r.t the invariant regions at each denoising timestep and constrain the generated content to be harmonized with the invariant region content. Experimental results verify that ReAtCo consistently improves the controllability of video diffusion editing and achieves superior video editing performance.

Code — https://github.com/mdswyz/ReAtCo
Extended version — https://arxiv.org/abs/2412.11710

Introduction

Text-guided video editing is a specialized facet of content creation, which can edit video content, including but not limited to manipulating objects, changing backgrounds, by manipulating the text prompt describing the source video. This task exemplifies the potential to augment and polish content within diverse domains, encompassing advertising design, marketing, and social media content (Zhao et al. 2023).

Recently, diffusion-based generative paradigm (Ho, Jain, and Abbeel 2020) has shown astonishing text-to-image (T2I) (Rombach et al. 2022; Saharia et al. 2022) and text-to-video (T2V) (Ho et al. 2022a; Blattmann et al. 2023) generation capabilities, which provides a great opportunity to manipulate video content via text guidance. To edit videos with low computational costs, some studies utilize large-scale pretrained T2I diffusion models, e.g., Stable Diffusion (Rombach et al. 2022) to develop various text-guided video editing methods (Wu et al. 2023a; Qi et al. 2023). The main idea of these methods is to flatten the temporal dimensionality of the source video and diffuse the flattened video into noise, and then the inverted noise is gradually denoised to the edited videos by the T2I-based video diffusion editing model under the condition of the edited text prompt. Moreover, due to the inherent absence of temporal awareness in T2I diffusion models, off-the-shelf methods tend to incorporate some additional modules or mechanisms to construct a well-designed video diffusion editing model, thus preserving the temporal consistency of edited videos. For example, Tune-A-Video (Wu et al. 2023a) incorporated the temporal attention modules and spatio-temporal attention modules into the T2I models for temporal coherence. FateZero (Qi et al. 2023) proposed a fusing attention mechanism to fuse the attention maps from the diffusion and

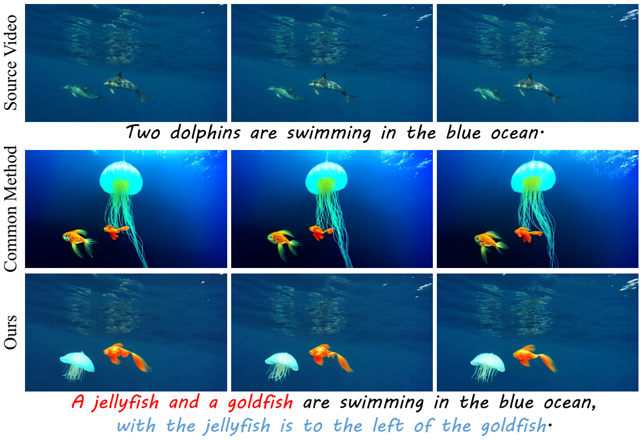

Source Video
Two dolphins are swimming in the blue ocean.
Common Method
Ours
A jellyfish and a goldfish are swimming in the blue ocean,
with the jellyfish is to the left of the goldfish.

Figure 1: Edited samples from the common video diffusion editing method (classic Tune-A-Video (Wu et al. 2023a) as an example) and our proposed ReAtCo.

[*]Corresponding authors: Xiaoya Zhang and Zhen Cui

generation process to facilitate motion consistency. TokenFlow (Geyer et al. 2024) designed a propagation mechanism to propagate a small set of edited features across frames.

Despite the great success, the controllability of editing remains a formidable challenge when performing fine-grained manipulation of multiple foreground objects. As shown in Fig. 1, the results from the common method show mislocated objects (i.e., the jellyfish is above the goldfish which is not aligned with *"the jellyfish is to the left of the goldfish"*) and incorrect number of objects (i.e., two goldfish and a jellyfish are generated which do not match *"A jellyfish and a goldfish"*). The essence behind this situation is the lack of spatial location awareness for the pretrained T2I models (Wu et al. 2023c,d). A question arises: *Can we improve the controllability of video editing based on off-the-shelf methods?*

In this paper, we aim to challenge the above limitations by proposing a Re-Attentional Controllable Video Diffusion Editing (ReAtCo) method. To efficiently control the spatial location of the edited objects aligned with the edited text prompts in a training-free manner, a Re-Attentional Diffusion (RAD) is proposed to refocus the cross-attention activation responses between the editied prompt and video content during the denoising stage, resulting in a spatially location-aligned and semantically high-fidelity target video. In addition, as each denoising timestep may lead to some sampling errors (Daras et al. 2024), the invariant region content that may exist during editing (e.g., the background in Fig. 1) is inevitably disrupted, ultimately resulting in a generated invariant region content that is far from the original ones. Therefore, we design an Invariant Region-guided Joint Sampling (IRJS) strategy to mitigate the sampling errors of the invariant region by injecting the original invariant region content into the denoising process, thus maintaining the invariant region information and constraining the generated content to be harmonized with the invariant region.

In contrast to prior works, our proposed ReAtCo could bring two benefits:**1)** ReAtCo can provide the ability for fine-grained manipulation of multiple foreground objects. As shown in Fig. 1, our ReAtCo can successfully edit *"two dolphins"* into *"a jellyfish and a goldfish"* while ensuring their spatial locations aligned with the target prompt (i.e., *"the jellyfish is to the left of the goldfish"*). **2)** the invariant region content could be faithfully preserved and the generated content is harmonized with the invariant region. We can observe from Fig. 1 that the background region (i.e., the invariant region in this case) content is consistently preserved while editing the two foreground objects. In summary, the contributions of this work can be concluded as:

- To improve the controllability of video editing, we propose a Re-Attentional Controllable Video Diffusion Editing (ReAtCo) method. ReAtCo can refocus the cross-attention activation responses by a well-designed RAD to control the spatial location of the edited objects aligned with the edited text prompt in a training-free manner.
- To keep the consistency of the invariant region with less border artifacts maximally, we design an IRJS to mitigate the sampling errors of the invariant region at each denoising timestep and to constrain the generated content to be harmonized with the invariant region.
- We perform extensive experiments and achieve superior or comparable results, demonstrating that our ReAtCo mitigates the limitations of existing state-of-the-arts, such as mislocated objects, incorrect number of objects.

Related Works

Text-to-image/video Generation. Text-to-image (T2I) generation task aims to generate photorealistic images that semantically match given text prompts (Mansimov et al. 2016; Ramesh et al. 2021; Rombach et al. 2022; Shen et al. 2024a). The main idea of this task is to utilize the generative models (Goodfellow et al. 2014; Ho, Jain, and Abbeel 2020; Wang, Cui, and Li 2023; Wang, Li, and Cui 2024) to construct a text-conditioned generative model with various attention or Transformer mechanism (Vaswani et al. 2017; Li et al. 2018; Zhang et al. 2020; Li, Zeng, and Shan 2020; Wang et al. 2022; Li and Shan 2023). Recently, due to powerful data generation capabilities, diffusion-based generative models have achieved great success in the T2I generation (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022; Luo et al. 2024; Shen et al. 2024b). For example, (Ramesh et al. 2022) proposed the DALLE-2 that uses CLIP-based (Radford et al. 2021) feature embedding to build a T2I diffusion model with improved text-image alignments. (Rombach et al. 2022) proposed a novel Latent Diffusion Model (LDM) paradigm that projects the original image space into the latent space of an autoencoder to improve T2I training efficiency. Despite the great success, text-to-video (T2V) generation is still extremely challenging due to the thousands of times harder to train compared to T2I models. Some researchers have attempted to challenge the T2V generation task and have proposed various methods (Ho et al. 2022a; Ge et al. 2023; Qing et al. 2023). For instance, (Ho et al. 2022b) proposed a Video Diffusion Model that uses a space-only 3D Unet to fit video content. (Blattmann et al. 2023) applied the LDM to high-resolution video generation.

Controllable Text-to-image/video Generation. Different from the above naive text-to-image/video methods, some studies aim to conduct controllable text-to-image/video generation by exploiting additional prior conditions (Wu et al. 2023d; Zhang et al. 2023; Shen and Tang 2024; Shen et al. 2024c). For example, (Zhang, Rao, and Agrawala 2023) proposed a ControlNet that appended additional conditions, such as Canny edges, depth maps, human poses, to provide diverse image generation capabilities. With this work, (Zhang et al. 2023) and (Chen et al. 2023) extended the ControlNet to the video generation domain, thereby achieving controllable T2V generation. (Yang et al. 2023) and (Phung, Ge, and Huang 2024) leveraged the bounding boxes to constrain the object generation. (Avrahami et al. 2023) utilized the segmentation maps to control the generation regions.

Text-guided Video Editing. The goal of text-guided video editing is to generate a new video derived from a given source video and an edited text prompt (Wu et al. 2023a; Qi et al. 2023; Chai et al. 2023; Geyer et al. 2024). Compared to earlier works such as (Kasten et al. 2021), this technology can reduce manual labor as the users only need to edit the text prompts describing the source videos. Before the diffusion-based era, (Bar-Tal et al. 2022) proposed a Text2Live to conduct text-driven video editing. The main idea of Text2Live is to utilize the layered neural atlas model (Kasten et al. 2021) to map source video into the image-based 2D atlas domain, thereby reducing the difficulty of video editing. In the era of diffusion models, (Chai et al. 2023) exploited the pretrained T2I diffusion models to edit 2D atlas images, but training the atlas models requires tremendous computational and time costs (7 $\sim$ 8 hours for training each video). Another effective paradigm is to flatten the temporal dimensionality of the source video and leverage DDIM (Song, Meng, and Ermon 2021) for the video-to-noise inversion, and then the inverted noise is gradually denoised to the edited videos by the pretrained T2I diffusion models. For example, (Wu et al. 2023a) proposed a Tune-A-Video that flattens the temporal dimensionality of the source video and then edits it frame-by-frame using the T2I model to generate the target video. Of these, the extra temporal attention modules are injected into the T2I model to preserve the temporal consistency. (Wang et al. 2024) designed a temporal Unet to guarantee comprehensive temporal modeling. TokenFlow (Geyer et al. 2024) designed a cross-frame propagation mechanism to enhance the temporal smoothness.

Method

Problem Description

Problem Let $\mathcal{V} = (v_1, v_2, \cdots, v_m)$ denotes a source video that contains m video frames. $\mathcal{P}$ and $\mathcal{P}'$ denote the source prompt describing $\mathcal{V}$ and the edited prompt provided by the users, respectively. The goal of text-guided video editing is to generate a new video $\mathcal{V}'$ from source video $\mathcal{V}$ under the condition of the edited prompt $\mathcal{P}'$. We illustrate an example:

- **Source:** an initial video with a prompt *"Two dolphins are swimming in the blue ocean."*
- **Target 1:** output a video to change *"Two dolphins"* as *"Two goldfishes"*.

Recent state-of-the-art methods can excellently achieve the goal by modifying the prompt based on the pretrained text-to-image (T2I) diffusion models, such as *"Two goldfishes are swimming in the blue ocean."* for Target 1. However, the fine-grained controllability of video editing remains a formidable challenge, e.g., to simply continue the above example (a failure for most existing methods):

- **Target 2:** output a video to fine-grained manipulate *"Two dolphins"* by editing *"the left dolphin as a jellyfish"* and *"the right dolphin as a goldfish"*.

The reason behind this failure is that the employed base models (i.e., the pretrained T2I models) are typically trained on simple text descriptions, not including fine-grained spatial location descriptions between different objects (Wu et al. 2023c,d). In other words, these methods often lack spatial location awareness for controllable video editing. A question arises: *Can we improve the fine-grained controllability of video editing with training-free mode?* It is not necessary to rebuild a new training dataset with information-enriched long text descriptions and retrain a new model due to high resource requirements.

Idea The edited video could be partitioned into two parts: changed parts and the remaining unchanged part (e.g., background, which we denote as the invariant region). For these changed parts, the users more focus on those objects of interest, which could be decided by the input prompts $\mathcal{P}'$ and $\mathcal{P}$. Suppose n objects need to be manipulated, denoted $\{O_i|_{i=1}^{n}\}$, the remaining part except objects is denoted O^-. To bridge the latent semantic information from new prompt $\mathcal{P}'$ to the video as well as keep spatial location awareness, we use text-video cross-attention maps (between text and denoised videos) to associate the objects of interest, denoted $\mathcal{A}_{O_i}(\mathcal{V}(t), \mathcal{P}')$ for object O_i, where $\mathcal{V}(t)$ is a noisy video at the t-th sampling step of the denoising process. For the unchanged part such as the background region, we expect to perform a diffusion-identical transformation $\mathcal{D}_{\mathrm{I}}$ to prevent the disruption of the unchanged region. Formally, our video sampling process (t to t–1 timestep) is defined as:

$$\mathcal{V}(t{-}1) \leftarrow F(\mathcal{D}_{\mathrm{R}}(\mathcal{V}(t), \{\mathcal{A}_{O_i}(\mathcal{V}(t),\mathcal{P}')|_{i=1}^{n}\},\mathcal{P}'), \mathcal{D}_{\mathrm{I}}(\mathcal{V}_{O^-}(t),\mathcal{P}')), \tag{1}$$

where $\mathcal{D}_{\mathrm{R}}$ is the diffusion editor w.r.t the changeable objects, F is an integration operation, and $\mathcal{V}_{O^-}(t)$ is the unchanged part of $\mathcal{V}(t)$. Accordingly, there are two questions that need to be solved:

- Spatial-aware diffusion editor $\mathcal{D}_{\mathrm{R}}$: the spatial alignment problem between object prompts and intermediate sampled video in a training-free manner. We propose a **Re-Attentional Diffusion (RAD)**.
- Diffusion-identical transformation $\mathcal{D}_{\mathrm{I}}$: recovery unchanged region with less border artifacts when integrating with new-generated object regions. We propose an **Invariant Region-guided Joint Sampling (IRJS)**.

Overview Framework

The framework of ReAtCo is illustrated in Fig. 2. Given a source video, we first utilize DDIM Inversion (Song, Meng, and Ermon 2021) for the video-to-noise inversion. Then, the inverted noise is gradually denoised to the edited video by an off-the-shelf video diffusion editing model. To achieve controllable video editing, the user needs to specify the region of interest according to their edited text prompt, e.g., the regions of two dolphins in the case of Fig. 2. Subsequently, the region of interest can be transformed into a set of binary masks, which are injected into the denoising stage to refocus the cross-attention activation responses by our proposed RAD, resulting in a spatially location-aligned and semantically high-fidelity edited video. In addition, we propose an IRJS to mitigate the sampling errors of the invariant region to maintain the original invariant content and allow the generated content to be harmonized with the invariant region.

Re-Attentional Diffusion

Reviewing the mainstream video diffusion editing models (Wu et al. 2023a; Qi et al. 2023; Chai et al. 2023; Wang et al. 2024), where the interaction between the textual semantic space and the pixel space occurs in the cross-attention layers of the pretrained T2I model such as Stable

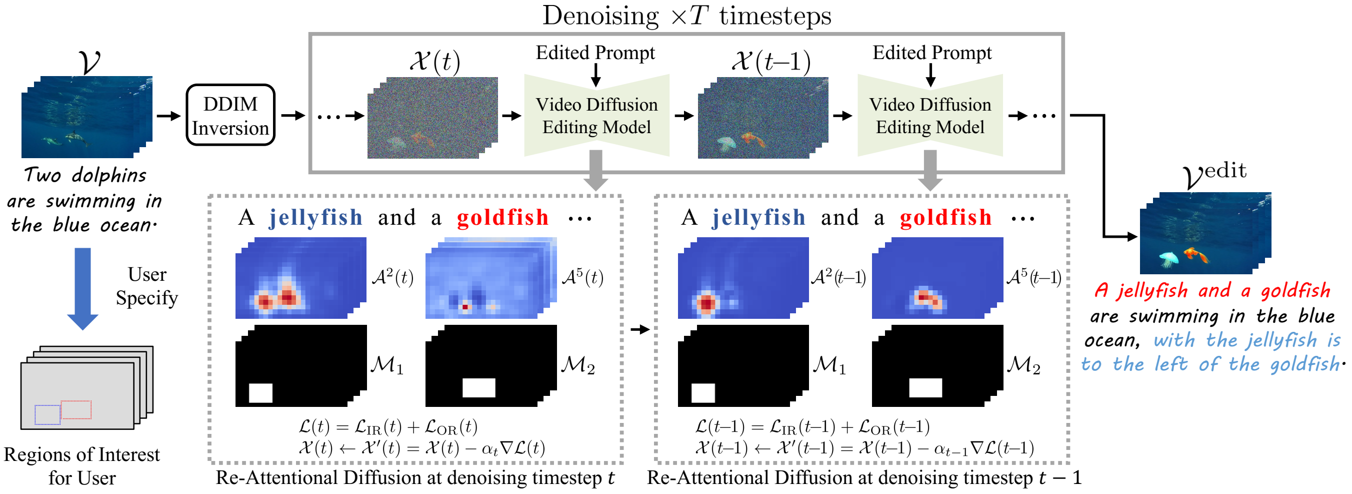

Denoising $\times T$ timesteps
$\mathcal{V}$
Two dolphins are swimming in the blue ocean.
DDIM Inversion
$\mathcal{X}(t)$
Edited Prompt
Video Diffusion Editing Model
$\mathcal{X}(t{-}1)$
Edited Prompt
Video Diffusion Editing Model
$\mathcal{V}^{\text{edit}}$
A jellyfish and a goldfish are swimming in the blue ocean, with the jellyfish is to the left of the goldfish.
User Specify
Regions of Interest for User
A jellyfish and a goldfish ···
$\mathcal{A}^2(t)$
$\mathcal{A}^5(t)$
$\mathcal{M}_1$
$\mathcal{M}_2$
$\mathcal{L}(t) = \mathcal{L}_{\text{IR}}(t) + \mathcal{L}_{\text{OR}}(t)$
$\mathcal{X}(t) \leftarrow \mathcal{X}'(t) = \mathcal{X}(t) - \alpha_t \nabla \mathcal{L}(t)$
Re-Attentional Diffusion at denoising timestep t
A jellyfish and a goldfish ···
$\mathcal{A}^2(t{-}1)$
$\mathcal{A}^5(t{-}1)$
$\mathcal{M}_1$
$\mathcal{M}_2$
$\mathcal{L}(t{-}1) = \mathcal{L}_{\text{IR}}(t{-}1) + \mathcal{L}_{\text{OR}}(t{-}1)$
$\mathcal{X}(t{-}1) \leftarrow \mathcal{X}'(t{-}1) = \mathcal{X}(t{-}1) - \alpha_{t-1} \nabla \mathcal{L}(t{-}1)$
Re-Attentional Diffusion at denoising timestep $t-1$

Figure 2: The framework of our proposed ReAtCo. Given a source video $\mathcal{V}$, ReAtCo first utilizes DDIM Inversion for video-to-noise inversion, and then the inverted noise is gradually denoised to an edited video $\mathcal{V}^{\text{edit}}$ by a video diffusion editing model. During the denoising stage, ReAtCo injects the proposed Re-Attentional Diffusion (RAD) and the user-specified regions of interest (i.e., the regions of two dolphins $\mathcal{M}_1$, $\mathcal{M}_2$) into video diffusion editing model to refocus the cross-attention maps (e.g., $\mathcal{A}^2(t)$ and $\mathcal{A}^5(t)$ for word index 2 and 5 at timestep t) between words of interest ("jellyfish" and "goldfish") and noisy video (e.g., $\mathcal{X}(t)$ at timestep t), thereby controlling the spatial location of the edited objects.

Diffusion (Rombach et al. 2022). That means that each video frame is computed the cross-attention maps with the text embedding, thereby bridging the relationship between text and video. Reviewing the computation of cross-attention maps, taking the i-th video frame as an example, and assume that we obtain the noisy video frame feature $\mathbf{X}_i(t)$ at denoising timestep t. $\mathbf{X}_i(t)$ is multiplied by the learnable parameter $\mathbf{W}_Q$ to obtain Query $\mathbf{Q}_i(t) = \mathbf{W}_Q\mathbf{X}_i(t) \in \mathbb{R}^{H\times W\times C}$, where H, W, and C indicate the height, width, and the channel dimensionality. The input word embedding $\mathbf{E}$ is multiplied by the learnable parameter $\mathbf{W}_K$ to generate Key $\mathbf{K} = \mathbf{W}_K\mathbf{E} \in \mathbb{R}^{L\times C}$, where L is the number of text tokens. With $\mathbf{Q}_i(t)$ and $\mathbf{K}$, the cross-attention maps $\mathbf{A}_i(t)$ of the i-th frame at denoising timestep t can be computed as:

$$\mathbf{A}_i(t) = \text{Softmax}(\mathbf{Q}_i(t)\mathbf{K}^\top/\sqrt{d}) \in \mathbb{R}^{L\times H\times W}. \quad (2)$$

$\mathbf{A}_i(t)$ is a tensor with the size of $L \times H \times W$, which means that each word is associated with a $H \times W$ pixel space cross-attention map, the values inside represent the relevance of the word to the pixel space. At a high level, the high response region in the cross-attention map associated with each word is equivalent to the region of generating word concept in the video frames, i.e., the higher the response, the more the word concept is being attended to in that region, and the content generated in that region is more aligned with word concept.

Inspired by the above phenomenon and facts, therefore, by modifying the pixel space cross-attention map corresponding to the word of interest in $\mathbf{A}_i(t)$, we could constrain the pixel region in which the word concept is generated. Taking Fig. 2 as an example, the user can first specify the two regions for the left and right dolphins from the source video (specify manually or automatically using the object detector), and then two regions can be transformed into two sets of binary masks $\mathcal{M}_1 = \{\mathbf{M}_1^1, \mathbf{M}_2^1, \cdots \mathbf{M}_m^1\}$ and $\mathcal{M}_2 = \{\mathbf{M}_1^2, \mathbf{M}_2^2, \cdots \mathbf{M}_m^2\}$. In this case, the ultimate goal is to edit the content of $\mathcal{M}_1$ to a jellyfish and the content of $\mathcal{M}_2$ to a goldfish. Thus, the words of interest are *"jellyfish"* and *"goldfish"* (the indexes of words are $\mathcal{I} = \{2, 5\}$), and we can modify the 2-nd and 5-th cross-attention maps along the L dimensionality of $\mathbf{A}_i(t)$ to maximize the attention response in the $\mathbf{M}_i^1$ and $\mathbf{M}_i^2$ regions, respectively. Once the cross-attention maps of all video frames are carefully modified, we can obtain a spatially location-aligned and semantically high-fidelity target video. For modifying cross-attention maps, a simple way is to modify all responses inside the object regions to 1 and outside the object regions to 0, but such a straightforward way may collapse the denoising process, potentially leading to a collapse of video fidelity.

Therefore, we propose a Re-Attentional Diffusion (RAD) that contains an inner-region of object constraint and an outer-region of object constraint, over the target cross-attention maps to gradually update the noisy video sample at arbitrary denoising timestep t such that the spatial location of edited objects will be aligned with the target regions.

Inner-Region of Object Constraint. To ensure the edited objects approach the user-specified regions, an intuitive objective is to ensure that high responses of cross-attention maps are in the target regions. Thus, we can build the inner-region of object constraint $\mathcal{L}_{\text{IR}}(t)$ at denoising timestep t:

$$\mathcal{L}_{\text{IR}}^j(t) = 1 - \frac{1}{K\times m}\sum_{i=1}^{m}\sum_{k=1}^{K}\text{top}_k(\mathbf{A}_i^j(t)\times\mathbf{M}_i^j, K), \quad (3)$$

$$\mathcal{L}_{\text{IR}}(t) = \sum_{j\in\mathcal{I}}\mathcal{L}_{\text{IR}}^j(t), \quad (4)$$

where $\mathcal{L}_{\text{IR}}^j(t)$ denotes the constraint corresponding to word index $j \in \mathcal{I}$. $\mathbf{A}_i^j(t)$ denotes the a cross-attention map corresponding to word index j in the i-th video frame at denoising timestep t, where $\mathbf{A}_i^j(t) \in \mathcal{A}^j(t)$ and $\mathcal{A}^j(t) = \{\mathbf{A}_1^j(t), \mathbf{A}_2^j(t), \cdots, \mathbf{A}_m^j(t)\}$ is a set of cross-attention maps for word index j in m video frames. $\mathbf{M}_i^j$ denotes the target region mask of the word concept corresponding to word index j in i-th video frame. $\text{top}_k(\cdot, K)$ represents that K elements with the highest response would be selected, which can reduce the sensitivity of the model to the masks (i.e., no precise masks are required). In the experiments, K is set as 20% of the number of the mask regions so that K is adaptively set according to the size of the mask.

Outer-Region of Object Constraint. The Inner-region of object constraint can control the edited object to appear inside the mask region, but it cannot ensure that the edited object is not synthesized outside the mask region. To mitigate the above issue, we further build a outer-region of object constraint $\mathcal{L}_{\text{OR}}(t)$ at denoising timestep t:

$$\mathcal{L}_{\text{OR}}^j(t) = \frac{1}{K \times m} \sum_{i=1}^{m} \sum_{k=1}^{K} \text{top}_k(\mathbf{A}_i^j(t) \times (1 - \mathbf{M}_i^j), K), \quad (5)$$

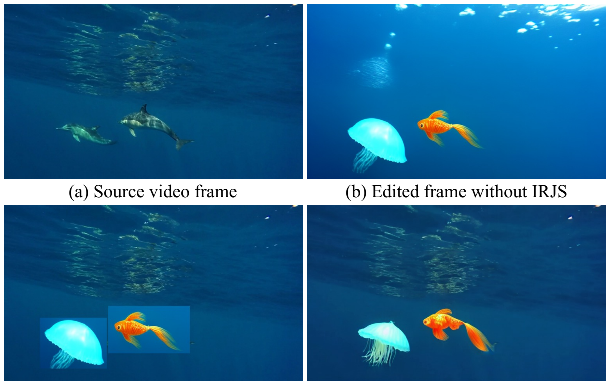
$$\mathcal{L}_{\text{OR}}(t) = \sum_{j \in \mathcal{I}} \mathcal{L}_{\text{OR}}^j(t). \quad (6)$$

Intuitively, $\mathcal{L}_{\text{OR}}(t)$ aims to minimize the activation responses of cross-attention maps out of the mask region, so that $\mathcal{L}_{\text{IR}}(t)$ and $\mathcal{L}_{\text{OR}}(t)$ constrain the cross-attention maps in a complementary manner.

Objective Optimization. We integrate the above constraints to reach the final RAD objective at denoising timestep t: $\mathcal{L}(t) = \mathcal{L}_{\text{IR}}(t) + \mathcal{L}_{\text{OR}}(t)$. Then, the noisy video sample $\mathcal{X}(t)$ could be updated with a step size of α_t as:

$$\mathcal{X}(t) \leftarrow \mathcal{X}'(t) = \mathcal{X}(t) - \alpha_t \nabla \mathcal{L}(t), \quad (7)$$

where α_t decays linearly at each denoising timestep. With the above constraints, $\mathcal{X}(t)$ at each timestep gradually moves toward the direction of generating high response attention in the given mask regions, thereby editing the target objects in the user-specified regions.


(a) Source video frame

(b) Edited frame without IRJS

(c) Edited frame with a direct copy of (a)

(d) Edited frame with IRJS

Figure 3: Edited video frames by different methods.

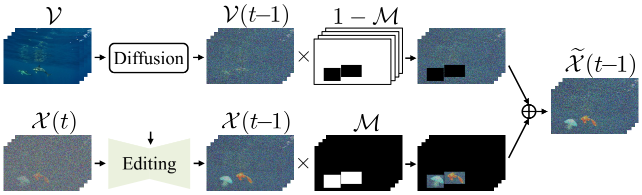


Figure 4: The framework of our proposed IRJS.

Invariant Region-guided Joint Sampling

The proposed RAD can refocus the cross-attention activation responses to control the editing region. However, we observe that when the user merely wants to edit foreground objects or edit partial foreground objects, e.g., in the case of Fig. 3, two dolphins need to be edited and the background region is the remaining invariant region, the generated invariant region content is often inconsistent with the original invariant region content. As shown in Fig. 3 (b), we can observe that although the edited frame is well-aligned with the edited prompt due to the nice property of RAD, the background region is inconsistent with the one of the source video frame (i.e., Fig. 3 (a)). This is because each denoising timestep leads to some sampling errors, and the accumulated errors from all timesteps eventually result in a generated background region that is far from the original background region. From the user's perspective, we would like to keep the original invariant background information when manipulating foreground objects. To preserve the content of the invariant region during the editing process, a straightforward idea is to copy the corresponding content from the source video directly into the target video, as shown in Fig. 3 (c). Intuitively, the object region is not harmonized with the background region, resulting in obvious border artifacts.

To mitigate the above issues, we propose an Invariant Region-guided Joint Sampling (IRJS) strategy to mitigate sampling errors of the invariant region by injecting the original invariant region content into the denoising stage and to constrain the generated content to be harmonized with the original invariant region content. The framework of IRJS is illustrated in Fig. 4, where we take the timestep t to $t-1$ as an example. For the vanilla sampling strategy (Ho, Jain, and Abbeel 2020), the noisy video sample $\mathcal{X}(t)$ at timestep t could be denoised into a noisy sample $\mathcal{X}(t-1)$ at timestep $t-1$ by a video diffusion editing model, but it may disrupt the information of the invariant region. The goal of IRJS is to mitigate the sampling error at each timestep by injecting the invariant region of the diffused source video sample into $\mathcal{X}(t-1)$. Specifically, the source video $\mathcal{V}$ is first diffused into a noisy sample $\mathcal{V}(t-1)$ at timestep $t-1$ according to predefined diffusion noise scheduler (Ho, Jain, and Abbeel 2020). Then, we use the object masks $\mathcal{M}$ (containing all object regions) and invariant region masks $1-\mathcal{M}$ to extract the object region of $\mathcal{X}(t-1)$ and invariant region of $\mathcal{V}(t-1)$, respectively. Finally, the extracted regions are added to obtain a noisy sample $\widetilde{\mathcal{X}}(t-1)$:

$$\widetilde{\mathcal{X}}(t-1) = \underbrace{\mathcal{X}(t-1) \times \mathcal{M}}_{\text{Generated Object Region}} + \underbrace{\mathcal{V}(t-1) \times (1-\mathcal{M})}_{\text{Original Invariant Region}}, \quad (8)$$

where $\mathcal{X}(t-1) \sim \mathcal{N}(\mu_\theta(\mathcal{X}(t), t), \Sigma_\theta(\mathcal{X}(t), t))$ and $\mathcal{V}(t-$

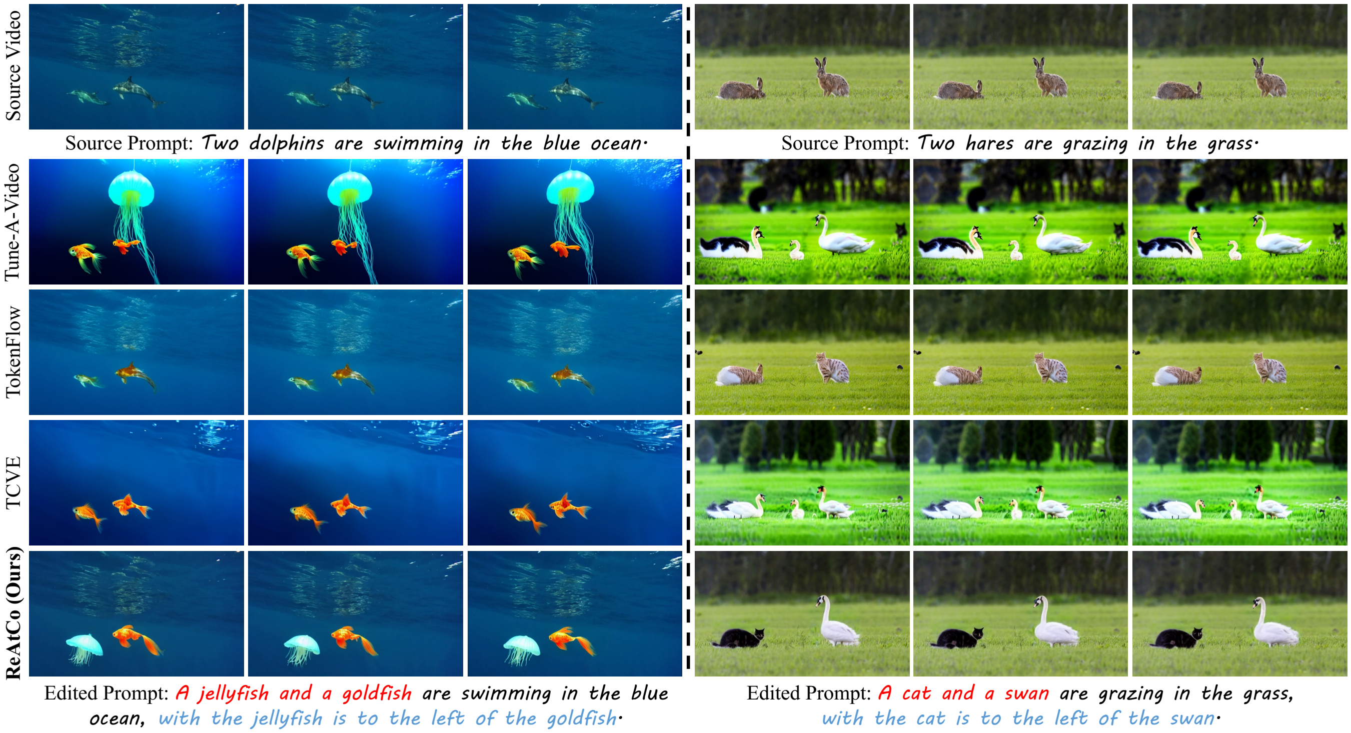

Source Video
Tune-A-Video
TokenFlow
TCVE
ReAtCo (Ours)
Source Prompt: Two dolphins are swimming in the blue ocean.
Source Prompt: Two hares are grazing in the grass.
Edited Prompt: A jellyfish and a goldfish are swimming in the blue ocean, with the jellyfish is to the left of the goldfish.
Edited Prompt: A cat and a swan are grazing in the grass, with the cat is to the left of the swan.

Figure 5: Visual comparisons of different methods in various scenes. Compared with these state-of-the-arts, ReAtCo can edit real-world videos with spatial location alignment, consistent number of objects, and high semantic fidelity.

$1) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathcal{V}(0), (1-\bar{\alpha}_t)\mathbf{I})$. Concretely, $\mu_\theta(\mathcal{X}(t), t)$ and $\Sigma_\theta(\mathcal{X}(t), t)$ are the predicted parameters of Gaussian transition distribution in the sampling (i.e., denoising) process, and $\bar{\alpha}_t$ is the total noise variance in the diffusion process predefined by (Ho, Jain, and Abbeel 2020). Further, when the video diffusion editing model is well-trained, then $\mathcal{N}(\mu_\theta(\mathcal{X}(t), t), \Sigma_\theta(\mathcal{X}(t), t)) \approx \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathcal{V}(0), (1-\bar{\alpha}_t)\mathbf{I})$. This is because the objective of the sampling process is to estimate the transition distribution of the diffusion process at each timestep. Thus, we can derive $\widetilde{\mathcal{X}}(t-1) \sim \mathcal{N}(\mu_\theta(\mathcal{X}(t), t), \Sigma_\theta(\mathcal{X}(t), t))$, abided by the distribution of $\mathcal{X}(t-1)$, so that we have $\mathcal{X}(t-1) \leftarrow \widetilde{\mathcal{X}}(t-1)$ that will be used as input for the next iteration in sampling process.

With multiple iterations of IRJS, the generated object region content could be harmonized with the original invariant region content. As shown in Fig. 3 (d), we can observe two benefits: **1)** the background region (i.e., invariant region) of the edited video is consistent with the source video. **2)** the object region is harmonized with the background region.

Experiments

Implementation Details

We conduct experiments on the text-guided video editing dataset LOVEU-TGVE-2023 (Wu et al. 2023b), the video samples used in (Chai et al. 2023), and the video samples from (Videvo 2024). Each video has 4 different edited prompts for evaluation. For specifying the object regions, we consider enabling the user to provide it in the possibly simplest way, i.e., bounding boxes. We consider three standard evaluation metrics that are proposed in the (Wu et al. 2023b) to measure the quality of edited videos. Frame Consistency is to measure the temporal consistency in frames by computing CLIP image embeddings on all frames of output video and reporting the average cosine similarity between all pairs of video frames. Textual Alignment is to measure the textual faithfulness of the edited video by computing the average CLIP score between all frames of the output video and the corresponding edited prompt. PickScore (Kirstain et al. 2023) is to measure human preference for T2I models. We compute the average PickScore in all frames of the output video. Furthermore, to measure the spatial location relationships between objects, we introduce the VISOR (Gokhale et al. 2022) that evaluates the spatial relationships (including left, right, above, below) in T2I generation. We compute the average VISOR in all frames of the output videos.

Baseline Comparisons

We compare our ReAtCo with the current state-of-the-arts, including the pioneer in efficient T2I-based video diffusion editing Tune-A-Video (Wu et al. 2023a), the fusing attention mechanism-based method FateZero (Qi et al. 2023), the atlas model-based method StableVideo (Chai et al. 2023), the dual-Unet architecture-based method TCVE (Wang et al. 2024), and the propagation mechanism-based method TokenFlow (Geyer et al. 2024). Below, we analyze quantitative and qualitative experiments.

Quantitative results. Tab. 1 lists the quantitative results

Methods	F-Consistency	T-Alignment	PickScore	VISOR
Tune-A-Video	92.54	26.75	20.37	15.62
FateZero	93.20	26.27	20.42	19.37
StableVideo	93.86	24.41	19.45	10.31
TokenFlow	94.66	26.89	20.57	11.56
TCVE	94.79	27.71	20.58	25.31
ReAtCo	**95.24**	**28.64**	**20.70**	**70.62**

Table 1: Quantitative comparison with evaluated baselines.

of different methods. From these results, we can observe that ReAtCo achieves the best video editing performance under four evaluation metrics. In particular, ReAtCo gains considerable performance improvements in the VISOR metric used to measure the spatial location relationships between objects. This observation could be ascribed to the fact that ReAtCo can control the spatial location of the edited objects by the well-designed RAD. Further analysis of the generated videos is provided in the next part.

Qualitative results. We showcase some visual comparison in Fig. 5. For the first sample, Tune-A-Video suffers from mislocated objects (i.e., the jellyfish is above the goldfish, which is unaligned with *"the jellyfish is to the left of the goldfish"*) and incorrect number of objects (i.e., a jellyfish and two goldfishes are generated, which is inconsistent with *"a jellyfish and a goldfish"*). TokenFlow and TCVE fail to edit the dolphin as the jellyfish. In contrast, ReAtCo can output a spatially location-aligned and semantically high-fidelity edited video. For the second sample, it is evident that only ReAtCo can faithfully modify the hare on the left to a cat and the hare on the right to a swan. The above phenomenon is attributed to our proposed RAD. At the same time, due to the benefit of IRJS, ReAtCo can also preserve the original background content and there are no obvious border artifacts between the foreground and background regions, showing better harmonization.

Ablation Studies

Quantitative analysis. We evaluate the effects of the key components in ReAtCo, including RAD and IRJS. The results are reported in Tab. 2, we conclude the conclusions as: **1)** Editing videos with RAD is effective, this is because RAD can empower the video diffusion editing model to perceive the spatial location of the foreground objects, thus improving the controllability and performance of video editing. Further, IRJS can bring some performance improvement by maintaining information in the invariant region and constraining the generated content to be harmonized with the invariant region. **2)** Combining RAD with IRJS brings further benefits, which proves that editing objects while maintaining invariant region content is feasible and effective.

RAD	IRJS	F-Consistency	T-Alignment	PickScore	VISOR
✓	✓	**95.24**	**28.64**	**20.70**	**70.62**
✓	×	94.59	28.54	20.64	69.06
×	✓	92.97	26.94	20.45	16.25
×	×	92.54	26.75	20.37	15.62

Table 2: Ablation study of the key components in ReAtCo.

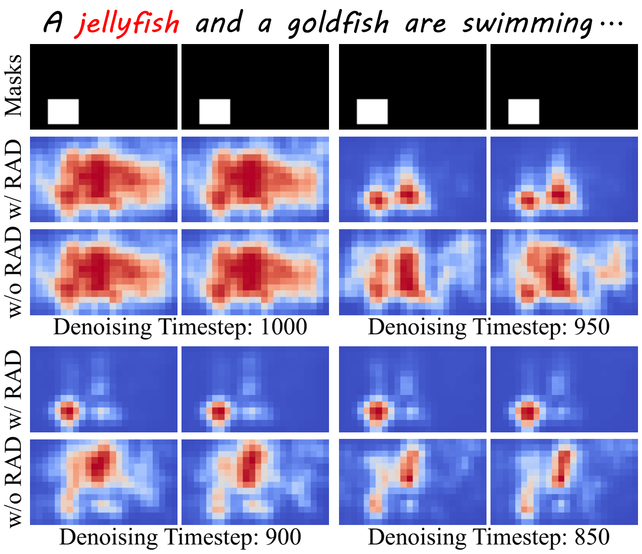


Figure 6: Visualization of cross-attention maps.

Visualization of cross-attention maps. We take the *"jellyfish"* in the first sample in Fig. 5 as an example to visualize the cross-attention maps during the denoising process. Fig. 6 shows the visualization of cross-attention maps associated with the word *"jellyfish"* from **w/ RAD** and **w/o RAD**, we can observe that the cross-attention responses in the initial denoising timestep (i.e., denoising timestep is 1000) are all in an irregular state. As the denoising timestep decreases, the cross-attention responses gradually focus on a region. In particular, for **w/o RAD**, the focused region of cross-attention responses gradually deviates from the user-specified region of the jellyfish. In contrast, cross-attention responses from **w/ RAD** gradually focus on the user-specified region of the jellyfish, which supports the effectiveness of RAD in refocusing cross-attention activation responses.

Conclusion

In this paper, we have proposed a Re-Attentional Controllable Video Diffusion Editing (ReAtCo) method for text-guided video editing. ReAtCo is inspired by the observation that the controllability of existing editing methods is not enough, especially in the controllability of spatial location. To efficiently improve the controllability of video editing, ReAtCo refocuses the cross-attention activation responses by the well-designed RAD to control the spatial location of the edited objects aligned with the edited text prompts in a training-free manner. In particular, we design an IRJS to preserve the invariant region information during editing and to constrain the generated content to be harmonized with the invariant region. Extensive experiments demonstrate the effectiveness of our ReAtCo.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants No. 62476133), the Research Grants Council of Hong Kong (Collaborative Research Fund

No. C7055-21GF) and by the Hong Kong Scholars Program, the Natural Science Foundation of Shandong Province (Grant No. ZR2022LZH003).

References

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18370–18380.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, 707–723. Springer.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chen, W.; Wu, J.; Xie, P.; Wu, H.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840*.

Daras, G.; Dagan, Y.; Dimakis, A.; and Daskalakis, C. 2024. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *Advances in Neural Information Processing Systems*, 36.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22930–22941.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2024. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *The Twelfth International Conference on Learning Representations*.

Gokhale, T.; Palangi, H.; Nushi, B.; Vineet, V.; Horvitz, E.; Kamar, E.; Baral, C.; and Yang, Y. 2022. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*.

Kasten, Y.; Ofri, D.; Wang, O.; and Dekel, T. 2021. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6): 1–12.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*.

Li, Y.; and Shan, S. 2023. Contrastive learning of person-independent representations for facial action unit detection. *IEEE Transactions on Image Processing*, 32: 3212–3225.

Li, Y.; Zeng, J.; and Shan, S. 2020. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 302–317.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2018. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 28(5): 2439–2450.

Luo, J.; Wang, Y.; Gu, Z.; Qiu, Y.; Yao, S.; Wang, F.; Xu, C.; Zhang, W.; Wang, D.; and Cui, Z. 2024. MMM-RS: A Multi-modal, Multi-GSD, Multi-scene Remote Sensing Dataset and Benchmark for Text-to-Image Generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mansimov, E.; Parisotto, E.; Ba, J. L.; and Salakhutdinov, R. 2016. Generating images from captions with attention. In *International Conference on Learning Representations*.

Phung, Q.; Ge, S.; and Huang, J.-B. 2024. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7932–7942.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Qing, Z.; Zhang, S.; Wang, J.; Wang, X.; Wei, Y.; Zhang, Y.; Gao, C.; and Sang, N. 2023. Hierarchical Spatio-temporal Decoupling for Text-to-Video Generation. *arXiv preprint arXiv:2312.04483*.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024a. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2024b. Boosting consistency in story visualization with rich-contextual conditional diffusion models. *arXiv preprint arXiv:2407.02482*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024c. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Videvo. 2024. Free stock video footage. https://www.videvo.net/. Accessed: 2024-12-23.

Wang, Y.; Cui, Z.; and Li, Y. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22025–22034.

Wang, Y.; Li, Y.; and Cui, Z. 2024. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Li, Y.; Zhang, X.; Liu, X.; Dai, A.; Chan, A. B.; and Cui, Z. 2024. Edit Temporal-Consistent Videos with Image Diffusion Model. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(12).

Wang, Y.; Lu, T.; Zhang, Y.; Wang, Z.; Jiang, J.; and Xiong, Z. 2022. FaceFormer: Aggregating global and local representation for face hallucination. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6): 2533–2545.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023a. 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*.

Wu, J. Z.; Li, X.; Gao, D.; Dong, Z.; Bai, J.; Singh, A.; Xiang, X.; Li, Y.; Huang, Z.; Sun, Y.; He, R.; Hu, F.; Hu, J.; Huang, H.; Zhu, H.; Cheng, X.; Tang, J.; Shou, M. Z.; Keutzer, K.; and Iandola, F. 2023b. CVPR 2023 Text Guided Video Editing Competition. arXiv:2310.16003.

Wu, Q.; Liu, Y.; Zhao, H.; Bui, T.; Lin, Z.; Zhang, Y.; and Chang, S. 2023c. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7766–7776.

Wu, W.; Li, Z.; He, Y.; Shou, M. Z.; Shen, C.; Cheng, L.; Li, Y.; Gao, T.; Zhang, D.; and Wang, Z. 2023d. Paragraph-to-Image Generation with Information-Enriched Diffusion Model. *arXiv preprint arXiv:2311.14284*.

Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14246–14255.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077*.

Zhao, M.; Wang, R.; Bao, F.; Li, C.; and Zhu, J. 2023. ControlVideo: Adding Conditional Control for One Shot Text-to-Video Editing. *arXiv preprint arXiv:2305.17098*.