RAVE: <u>Ra</u>ndomized Noise Shuffling for Fast and Consistent <u>V</u>ideo <u>E</u>diting with Diffusion Models

Ozgur Kara[1*]   Bariscan Kurtkaya[2*†]   Hidir Yesiltepe[4]   James M. Rehg[1,3]   Pinar Yanardag[4]

[1]Georgia Tech   [2]KUIS AI Center   [3]UIUC   [4]Virginia Tech

okara7@gatech.edu, bkurtkaya23@ku.edu.tr, hidir@vt.edu, jrehg@uiuc.edu, pinary@vt.edu

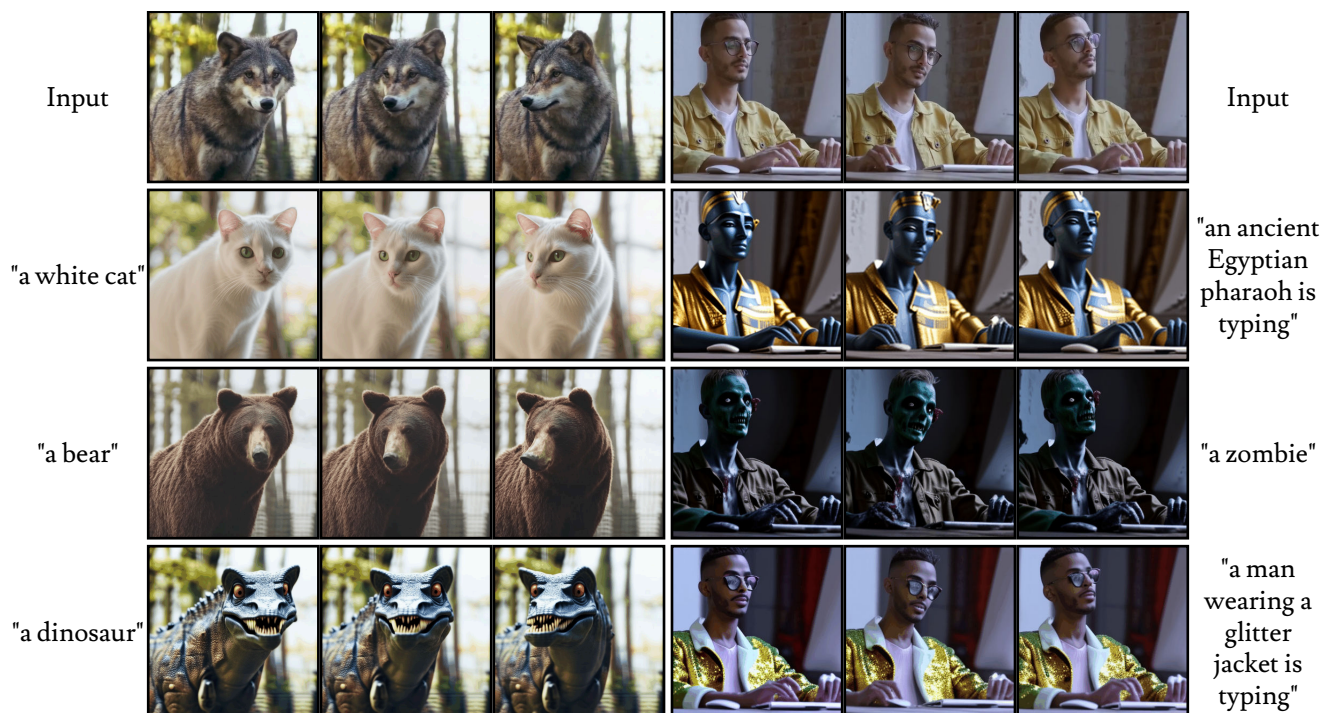Project Webpage: https://rave-video-edit.github.io

Figure 1. **RAVE** is a lightweight and fast video editing method that enhances temporal consistency in video edits, utilizing pre-trained text-to-image diffusion models. It is capable of modifying local attributes, like changing a person's *jacket* (bottom right), and can also handle complex shape transformations, such as turning a *wolf* into a *dinosaur* (bottom left).

Abstract

Recent advancements in diffusion-based models have demonstrated significant success in generating images from text. However, video editing models have not yet reached the same level of visual quality and user control. To address this, we introduce RAVE, a zero-shot video editing method that leverages pre-trained text-to-image diffusion models without additional training. RAVE takes an input video and a text prompt to produce high-quality videos while preserving the original motion and semantic structure. It employs a novel noise shuffling strategy, leveraging spatio-temporal interactions between frames, to produce temporally consistent videos faster than existing methods. It is also efficient in terms of memory requirements, allowing it to handle longer videos. RAVE is capable of a wide range of edits, from local attribute modifications to shape transformations. In order to demonstrate the versatility of RAVE, we create a comprehensive video evaluation dataset ranging from object-focused scenes to complex human activities like dancing and typing, and dynamic scenes featuring swimming fish and boats. Our qualitative and quantitative experiments highlight the effectiveness of RAVE in diverse video editing scenarios compared to existing methods. Our code, dataset

---

* Joint co-author
† B. Kurtkaya worked on this project as an intern at Virginia Tech.

and videos can be found in our project webpage.

1. Introduction

Diffusion-based generative models [16, 43] demonstrate outstanding capabilities in generating and editing diverse and high-quality images [36, 37, 40] guided by text prompts. Existing diffusion-based editing methods [6, 7, 12, 15, 29, 32, 39, 46] have predominantly addressed image-related tasks, and offer a wide range of functionalities such as object editing [15], image inpainting [6], personalized generation [39], and image-to-image translation [29]. These methods perform editing on images through a deterministic DDIM inversion process [44]. This involves conducting image-to-noise inversion and subsequently generating edited images from the noise based on the target text prompt. This new paradigm of interactive image editing and generation through textual prompts not only represents an exciting advancement but also lays the groundwork for a diverse spectrum of possibilities for content creation by nontechnical users.

Unfortunately, advances in image editing have not been translated into comparable success in video editing. Prior research has been focused on text-to-video (T2V) generation [17, 18, 42] by utilizing extensive text-video datasets such as WebVid-10M [8]. However, these methods are prohibitively expensive and therefore unsuitable for broad use in content creation. Another line of research employs conventional video editing methods [9, 23, 26] such as keyframe selection [22] or atlas editing [9, 23]. However, these methods are hampered by their dependence on time consuming processes like atlas learning [9, 23, 26], accurate keyframe selection [22], and prompt-dependent tuning [9, 26]. More recent works have focused on leveraging pre-trained text-to-image (T2I) diffusion models [9, 10, 34, 50, 51]. However, adapting pre-trained T2I models for video editing poses a significant challenge, as it is essential to ensure the temporal consistency within the edited video. While existing methods exhibit potential, they face multiple limitations: difficulty in editing long videos [10, 34], challenges with complex edits such as shape [14, 25, 51], quality degradation in the presence of substantial motion [25], demands significant processing time [34, 50], or requires additional training [27, 50]. In this paper, we present RAVE, a novel text-guided zero-shot video editing approach that performs style, attribute, and shape editing on videos. It is compatible with any pre-trained T2I diffusion model, such as Stable Diffusion (SD) [37], eliminating the need for training for every target prompt [9, 23, 26], or user-specific masks [6, 7], and also incorporates spatial guidance through ControlNet [52]. Notably, RAVE is lightweight and efficient, achieving video edits at a rate ∼25% faster than existing methods.
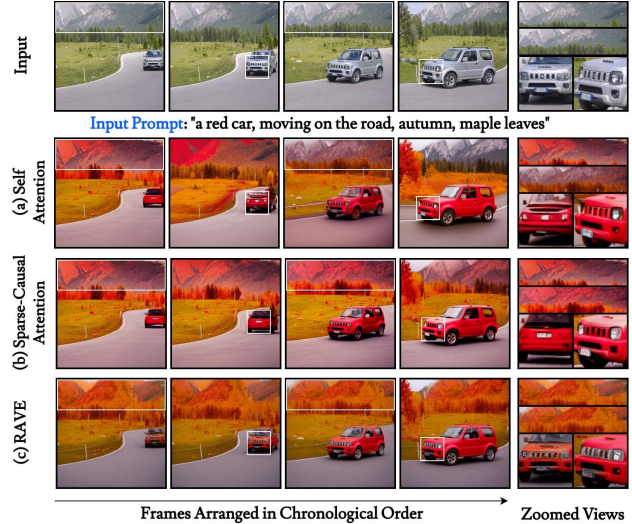

Figure 2. ***Comparison with existing attention modules.*** The first row shows the frames of the input video, followed by subsequent rows presenting the edited frames under various attention settings, with our approach in the last row. The column on the right provides zoomed-in crops of mountains (from the 1st and 3rd frames) and the car's bumper (from the 2nd and 4th frames).

Maintaining temporal consistency is crucial in video editing to ensure consistency across frames and to preserve the seamless movement of objects. Unlike previous studies [10, 34, 50] that depend on explicitly using multiple frames with attention methods like spatio-temporal attention or sparse causal attention for consistency [10, 50], RAVE employs a novel noise shuffling strategy that efficiently utilizes spatio-temporal interactions between frames. Our noise shuffling strategy inherently directs the model to perform spatio-temporal attention in a fraction of the usual time and memory requirements while still maintaining temporal consistency. Fig. 2 illustrates common attention strategies in video editing methods, highlighting their shortcomings and emphasizing the effectiveness of our noise shuffling approach. Using solely spatial self-attention can leverage pixel locations in feature maps to capture similar correlations, while cross-attention manages the correspondence between pixels and text prompts (see Fig. 2 (a)). However, as shown by the results, while the generated frames align with the text prompt in terms of motion and color style, they lack consistency due to the neglect of temporal context as seen in the inconsistencies in the background and the car's bumper. A straightforward approach would be to expand the self-attention module to include *all* frames, but this presents a substantial computational challenge [50]. Processing all video frames simultaneously becomes impractical, especially for videos with a high frame count. Alternative spatio-temporal attention methods often use a sparse form of the mechanism, known as *sparse-causal attention* [50] where the attention matrices are calculated between the initial frame and preceding frames, or among se-

lected keyframes [10, 19, 34, 50]. Although this method produces more consistent frames with reduced time complexity, its performance tends to decline in longer videos due to the diminishing temporal awareness. This decline is evident in Fig. 2 (b), as shown by the changes in the hood and the contrast in the mountains. In contrast, our strategy leverages the power of global spatio-temporal attention throughout *all* frames, ensuring frame consistency even in long videos by performing a novel noise-shuffling operation at every diffusion step during sampling (see Fig. 2 (c)). This results in maintaining the reduced time complexity when editing long videos without sacrificing the quality or speed of image editing. To summarize, our main contributions are as follows:

- We present RAVE, a novel, training-free, zero-shot video editing framework that integrates with pre-trained T2I diffusion models, such as Stable Diffusion [37]. RAVE incorporates spatial guidance with ControlNet [52] to perform video style, attribute, and shape editing.
- Our novel noise shuffling method harnesses spatio-temporal interactions between frames efficiently, resulting in temporally consistent videos ~25% faster than existing methods.
- We propose a novel video evaluation dataset that includes a diverse range of videos, from object-centric scenes to complex human motions like dancing and typing, as well as dynamic scenes featuring swimming fish and boats. Moreover, we share our source code publicly, hoping to promote further research.

2. Related Works

Text-driven image editing Methods such as Dream-Booth [39] and Textual Inversion [13] demonstrate diverse image generation through fine-tuning in a few-shot manner. UniTune [47] and Imagic [24], both based on the Imagen [40] model, exhibit strong editing performances. Recent training-free methods like Prompt-to-Prompt [15], DiffEdit [12], Blended Diffusion [6], and Blended Latent Diffusion [7] achieve local and detailed editing by leveraging attention properties.

Text-driven video editing with training Text-guided video editing broadly falls into two categories. One recent work in the first category, Video Diffusion Models [18], uses a factorized space-time U-Net compatible with joint image-video training as its architecture. Dreamix [31] extends this with a mixed fine-tuning approach. In contrast, Make-A-Video [42] reduces T2V training cost by eliminating the need for paired text-video data. MagicEdit [27] introduces a different learning setting, separating the learning of structure, motion, and content. However, these methods require extensive training on a large dataset, limiting scalability and generalizability. Tune-A-Video [50] proposes tailored spatio-temporal attention as an extension to the T2I model, termed sparse causal attention. Shape-aware TLVE [26] suggests an atlas-learning-based approach enabling structure-editing. Edit-A-Video [41] introduces 'sparse-causal blending' to reduce background inconsistency, utilizing Null-Text Inversion [30], akin to Video-P2P [28] and vid2vid-zero [48]. However, these methods are time-consuming, requiring optimization, which limits their real-world application capabilities.

Text-driven video editing without training Due to the optimization requirements of previous methods, recent studies emphasize zero-shot training-free approaches for practical applicability. Pix2Video [10] employs sparse-causal attention for temporal consistency along with latent guidance, using the predictions of the original images as a proxy at each denoising step. FateZero [34] uses attention features during inversion for spatio-temporal preservation and blending, claiming that those are better in preserving motion and structure compared to that of during sampling. Text2Video-Zero [25] synthesizes and edits videos with cross-frame attention, initial frame integration, and background smoothing. Rerender-A-Video [51] employs hierarchical cross-frame constraints for temporal consistency, while TokenFlow [14] focuses on feature-level smoothing to reduce the effects of flickering.

Nevertheless, FateZero [34], similar to Pix2Video [10], requires source prompts as it is built on the Prompt-to-Prompt [15] editing technique. This method necessitates specific types of prompts on the source prompt, thereby limiting editing diversity. Additionally, both are constrained to shorter clips due to memory limitations. Text2Video-Zero [25] and Rerender [51] heavily rely on off-the-shelf methods and optical flow, limiting consistency over longer videos. TokenFlow [14] faces structure editing limitations, dependent on inversion quality, and feature-level smoothing causes blurring in the edited image. As opposed to our method, previous approaches require time-intensive operations that slow down inference speed, restricting applicability to longer videos.

3. Methodology

Consider a video, denoted as $\mathcal{V}_K = \{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$, consisting of K individual frames. Each frame, represented as $\mathcal{I}_k \in \mathbb{R}^{W \times H \times C}$ corresponds to the kth position in the video sequence where W, H, and C specify the width, height, and color channels, respectively. Suppose we have a text prompt $\mathcal{P}$, detailing a specific edit to be made to the video. Our goal is to generate a modified video, $\mathcal{V}_K^*$, where the original video is transformed based on the changes outlined in $\mathcal{P}$. For example, imagine a video of a monkey sitting on the grass and eating food (refer to Fig. 6). If given a prompt such as *"a teddy bear is eating an apple"*, our task would be to transform the monkey and food into a teddy bear and apple, while preserving its original motion and
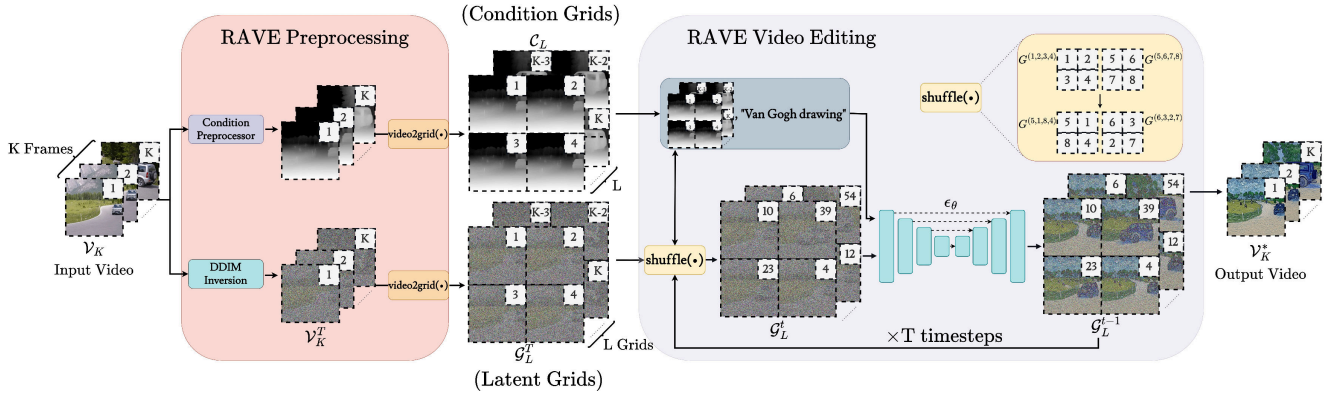
Figure 3. *An illustration of RAVE.* Our process begins by performing a DDIM inversion with the pre-trained T2I model and condition extraction with an off-the-shelf condition preprocessor applied to the input video ($\mathcal{V}_K$). These conditions are subsequently input into ControlNet. In the RAVE video editing process, diffusion denoising is performed for T timesteps using condition grids ($\mathcal{C}_L$), latent grids ($\mathcal{G}_L^t$), and the target text prompt as input for ControlNet. Random shuffling is applied to the latent grids ($\mathcal{G}_L^t$) and condition grids ($\mathcal{C}_L$) at each denoising step. After T timesteps, the latent grids are rearranged, and the final output video ($\mathcal{V}_K^*$) is obtained.

semantic layout. As we achieve this transformation using only publicly available T2I models combined with a single video-text pair, we consider our approach as *zero-shot text-guided video editing*.

3.1. Preliminaries

Latent diffusion models (LDMs) Denoising diffusion probabilistic models (DDPM) are computationally intensive due to repetitive denoising steps in pixel space. Addressing this, LDMs [37] introduce two key extensions: (i) The diffusion process operates in a lower-dimensional space by initially encoding the image using a pre-trained autoencoder, and (ii) improved controllability is achieved through conditions during training. LDMs typically employ a U-Net [38] as the denoising model, featuring skip connections between decoder and encoder levels. The U-Net consists of stacked 2D convolutional residual blocks and transformer blocks, each with a spatial *self-attention* (S-A) layer capturing spatial correlations, a *cross-attention* (C-A) layer interacting with conditional inputs (*e.g.* tokens in the text prompt), and a *feed-forward layer* processing the output for compatibility with the subsequent block.

Denoising diffusion implicit models (DDIM) To expedite the sampling process of diffusion models, DDIM [44] generalizes the Markov structure of DDPM [16] to a non-Markov setting and allows us to reach the clear sample with less steps. That allows us to obtain the corresponding latent from the original latent, z_0, for each timestep t, $\{z_t\}_{t=1}^{t=T}$. For image editing, one needs to map the given image to its inverse counterpart. Therefore, this technique is widely utilized in various image/video editing approaches with diffusion models.

ControlNet To enhance flexibility and controllability of the video editing process, our method incorporates ControlNet [52] for structural guidance. ControlNet introduces

a control mechanism for diffusion models by integrating a weight locking strategy, allowing for additional conditions beyond the text prompt $\mathcal{P}$, such as line art, depth, pose, and surface normals. This feature facilitates more controlled generation, particularly in the structure of images, offering an advantage over traditional LDMs.

3.2. Our approach

Grid trick The grid trick, also known as *character sheet*, is a popular hack in SD [37] art community. It is frequently used by enthusiasts and hackers for avatar design and image stylization [2]. This approach allows for applying a consistent style to all images in a grid using a T2I model, as shown in Fig. S.9. The effectiveness of this simple method in producing multiple images with consistent style lies in how the diffusion model performs editing of a single grid. It treats all frames collectively as a single image through its convolution-based residual blocks and transformer blocks. The self-attention layers in the transformer blocks are especially crucial for enhancing temporal consistency across frames in a grid. Simultaneously, the convolution layers within the residual block are responsible for capturing both temporal connections and spatial correlations. Note that, to the best of our knowledge, we are the first to systematically utilize this trick in image/video editing research.

Grid trick for video editing A straightforward adaptation of this trick for video editing involves transforming the input video into a grid layout and then feeding this grid to T2I editing tools such as ControlNet. Subsequently, the modified grid can be unfolded into separate frames and sequenced consecutively to produce the final video. More formally, assume that each frame in the video is arranged into a $N = n \times m$ grid with n rows and m columns, denoted as $G^{(1,\cdots,N)} \in \mathbb{R}^{(W \cdot m) \times (H \cdot n) \times 3}$, where H and W represent the height and width of the frames in the video, respectively,
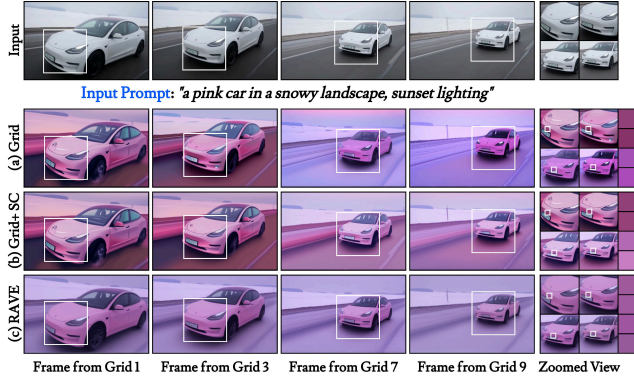
Figure 4. ***Consistency across grids.*** Editing results are shown for (a) processing grids independently, (b) adapting sparse-causal attention using grids, and (c) applying RAVE. The rightmost column features a close-up of the car's front, highlighting temporal color changes per approach. RAVE produces consistent patches in all grids while other methods struggle with consistency.

and superscript $(1, \cdots , N)$ denotes the indices of frames included in the grid. During the editing process of a single grid $G^{(1, \cdots , N)}$, the diffusion model processes N frames holistically as a single image by fostering a *smooth transition* between the latent vectors, $\{z_t^0, \ldots, z_t^n\}_{t=1}^{t=T}$, z_t^i denoting the corresponding latent region of ith frame within the grid at timestep t. This method achieves consistent frames in short sequences as can be seen from Fig. 4 (a), where the frames from grid 1 and grid 3 share the same style. However, it encounters challenges in maintaining this consistency in longer videos, even with the same text prompt. This limitation arises due to GPU memory constraints, restricting the number of frames that can fit within a grid. For example, if a GPU can only process a maximum of a 3×3 grid, equivalent to 9 frames, videos exceeding 9 frames need to be divided into multiple grids, each processed separately. Consequently, the sampling trajectories for each grid diverge, leading to internal consistency within each grid but noticeable style differences between grids, as evident in Fig. 4 (a), where the style of frames from separate grids differs.

Noise shuffling While the grid technique enables consistent editing, ensuring consistency *across* multiple grids remains a challenge. One could modify well-known attention mechanisms, like sparse-causal attention, for the grid structure. In this adaptation, attention is shifted from focusing on the initial frame and the previous frame to the *initial grid* and the *previous grid*. However, this approach can still face difficulties in maintaining consistency with longer videos. For instance, Fig. 4 (b) demonstrates that although sparse-causal attention works effectively for frames from grid 1 and grid 3, it encounters challenges in maintaining consistency as the video progresses. In order to achieve consistency across grids, we introduce a novel *noise shuffling* approach that ensures consistent styling across independent grids (see Fig. 4 (c)). Consider the *video2grid*$(\cdot, \cdot)$ function

that accepts an input video and a grid size $N = n \times m$, and generates a series of grids while preserving the original frame order. For instance, when the original video $\mathcal{V}_K$ is used as input, we can produce a set of L grids as:

$$video2grid(\mathcal{V}_K, N) = \{G_1^{(1, \cdots , N)}, \ldots, G_L^{(P, \cdots , K)}\} \quad (1)$$
$$= \mathcal{G}_L \quad (2)$$

where $P = K - (N + 1)$. In this context, L represents the total number of grids, calculated as $L = K/N$, assuming that K is divisible by N.

Rather than processing each grid in $\mathcal{G}_L$ sequentially to produce the final video output, which gives undesirable results (as depicted in Fig. 4 (a)), our approach entails randomly shuffling the order of frames at each timestep of the sampling process and allocating frames to different grids in a random order. More formally, starting with the input grids generated by the *video2grid*$(\cdot, \cdot)$ operation, we proceed as follows:

$$shuffle(\mathcal{G}_L) = \{G_1^{(i, \cdots , j)}, \ldots, G_L^{(k, \cdots , l)}\} \quad (3)$$

where $(i, \cdots , j), \cdots , (k, \cdots , l) \in \{1, \cdots , K\}$ are randomly chosen non-repeating indices. In other words, this function randomly rearranges the locations of the frames in a given set of grids. Through the random formation of grids from latents arranged in a random order, we encourage spatio-temporal interaction between each frame throughout the denoising process. This interaction involves convolutional layers, which play a role in globally smoothing the latent vectors, thereby reducing pixel shifts that can lead to flickering. Furthermore, self-attention layers ensure that features of each patch from every frame are influenced by patches of other frames at each step, guaranteeing the application of the same style globally. This enhancement contributes to temporal consistency, regardless of the video size since our approach does not require extra memory requirement proportional to the length of the video. The complete structure of our approach is depicted in Fig. 3.

4. Experimentation

Baselines We conduct a thorough quantitative and qualitative comparison with concurrent baselines, using their official implementations. Rerender [51] relies on optical flow, Text2Video-Zero [25] employs random warping and optional background smoothing, limiting its effectiveness in videos with background motions. TokenFlow [14] applies feature-level smoothing but has constraints in shape-aware editing. FateZero [34] is excluded from quantitative comparison due to its requirement for target prompts to be in a specific format with the source prompts, necessitating manually crafted prompts. Additionally, both FateZero [34] and Pix2Video [10] face GPU usage limits, which cannot process videos for more than 22 and 45 frames for Pix2Video

Figure 5. ***Qualitative comparison.*** We present a comparison of approaches on videos featuring diverse motions and objects on short and long videos. Videos in columns (a) and (b) have a total of 36 frames, while video in (c) has 90 frames. Due to their excessive GPU requirements, Pix2Video and FateZero are executed on shorter clips. Results are best viewed in zoomed-in. For more, please see our project webpage.

and FateZero with their official repositories, respectively. We use official repositories and default parameters of the baselines and the suggested parameters provided by the authors after reaching out to them and default parameters otherwise. We report the results of RAVE and RAVE without shuffling as an ablation. Please refer to the details of the dataset used in the experiments in SM S.1.

4.1. Implementation details

We use Stable Diffusion (SD) 1.5 from the official Huggingface repo for all baselines in any type of comparison. For qualitative assessment of our approach, the Realistic Vision V5.1 model from the CivitAI [1] repo is employed. All approaches use 50 steps of DDIM inversion and sampling along with a fixed classifier-free guidance scale of 7.5. RAVE employs a 2×2 grid size for videos with a length of 8 frames and a 3×3 grid size for the rest. In all comparisons, including quantitative, qualitative and user study comparison, we applied depth-conditioned ControlNet for our method as well as for Pix2Video, Text2Video-Zero, and Rerender. TokenFlow utilizes Plug-and-Play [46] for image editing method, therefore we use Plug-and-Play in their method during comparison. We do not use any negative or positive prompts in RAVE. Also for a fair comparison, no positive prompts or negative prompts are used for other methods as well. We run all experiments on a single A40 GPU.

4.2. Evaluation

Qualitative assessment Fig. 1 and Fig. 6 illustrate some examples of video editing using RAVE. We demonstrate style editing (*e.g.* watercolor style), complex shape editing (*e.g.* transforming a car into a tractor), and both applied simultaneously (*e.g.* 'a jeep is moving in the grassy field'). We use videos featuring backgrounds in motion (*e.g.* top right video) and different objects (*e.g.* turtle, human, watchtower) engaged in various activities (*e.g.* swimming, eating, stretching). Moreover, we present a qualitative comparison in Fig. 5 with baseline methods. Notice that FateZero struggles to edit non-local attributes, as evident with twinkling stars in Fig. 5 (b) since the quality of edit depends on Prompt-to-Prompt [15] which heavily relies on the chosen parameters and source prompt. Text2Video-Zero appears effective with videos featuring constant backgrounds, despite a color change in the woman's jacket Fig. 5 (c); however, it encounters challenges when there is motion in the background, such as the trees in Fig. 5 (a). Rerender heavily utilizes optical flow along with keyframe propagation, however these are not enough to perform editing while keeping style over time (observe the color change of the car in Fig. 5 (a)). Although TokenFlow excels at maintaining structural consistency over time, supported empirically by quantitative results (WarpSSIM), it experiences overall blurring in the entire image due to 'feature-level smoothing' being applied. Pix2Video's edits deviate significantly
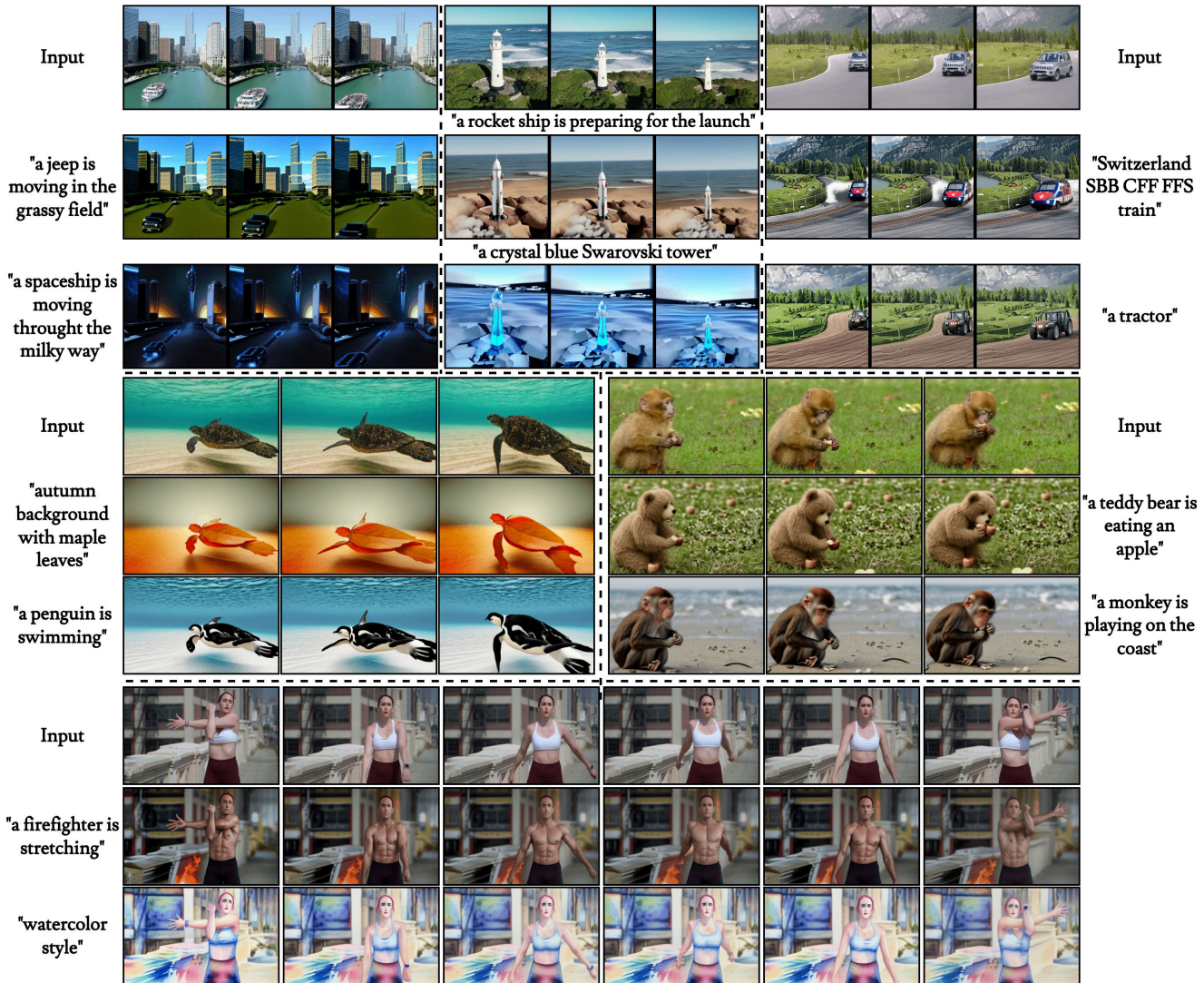
Figure 6. *Qualitative results with RAVE.* Our method is capable of performing a variety of video editing tasks, including style editing, local attribute editing, and shape editing (top right). It can edit videos featuring moving objects, such as turtles or boats, as well as videos with complex human movements, like a woman stretching (bottom row). Best viewed in zoomed-in. Full videos are provided in SM.

from the source video (note the woman in Fig. 5 (c) and the bear in Fig. 5 (b)), indicating that latent guidance alone and depth guidance solely are insufficient for consistency in long videos. In contrast, our method preserves global consistency, aligns with the text prompt, and maintains editing quality—all accomplished in less time compared to previous approaches (see Table 1 last column).

Quantitative evaluation Following prior work [10, 14, 51], we assess the temporal consistency of editing at a coarse level by measuring the average pairwise similarities (**CLIP-F**) of each frame using CLIP [35] and structural consistency (**WarpSSIM**) by calculating the average SSIM [49] score between the warped edited video (RAFT [45] is used to obtain the optical flow of the source video) and the edited video. Additionally, we evaluate the textual alignment of the editing, given a prompt, by calcu-

lating the average distance between the CLIP embedding of the text prompt and all frames (**CLIP-T**). Note that these metrics alone do not provide a comprehensive evaluation since CLIP-T is unrelated to consistency, and WarpSSIM and CLIP-F can yield high scores even when the video is not edited at all. Therefore, following [11], we adopt $\mathbf{Q_{edit}}$ = WarpSSIM · CLIP-T for a more holistic assessment. The results are shown in Table 1. Quantitative evaluation is performed on our dataset, consisting of 10, 15, and 6 videos with lengths of 8, 36, and 90 frames, respectively. Each video is edited with 4 style and 2 shape prompts, resulting in 186 text-video pairs. Our performance surpasses previous approaches in both frame consistency and textual alignment. While TokenFlow demonstrates slightly greater effectiveness for shorter frames in terms of structural consistency (WarpSSIM), our superiority becomes evident for

Table 1. ***Quantitative comparison.*** CLIP-F, WarpSSIM, CLIP-T, and Q_{edit} metrics are reported individually on videos of 8, 36, and 90 frames. The user study section reports the frequency of each method chosen among the top two edits for General Editing (Q1 (GE)), Temporal Consistency (Q2 (TC)), and Textual Alignment (Q3 (TA)). The last column presents video-editing runtime in 'minutes:seconds' format for 90 frames for the entire pipeline, including preprocessing and editing stages (parentheses indicate runtime w/o preprocessing). '-' denotes methods that cannot be measured due to excessive memory requirements, while 'N/A' indicates that the value is not available.

Method	CLIP-F ($\times 10^{-2}$)↑			WarpSSIM ($\times 10^{-2}$)↑			CLIP-T ($\times 10^{-2}$)↑			Q_{edit} ($\times 10^{-5}$)↑			User Study↑			Runtime↓
	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	Q1 (GE)	Q2 (TC)	Q3 (TA)	90-frames
Text2Video-Zero	95.49	92.89	94.35	67.97	36.65	71.57	29.46	29.42	29.73	20.02	10.78	21.27	47.95%	24.87%	52.56%	5:33
Rerender	92.87	89.71	90.63	68.57	44.54	74.56	25.65	27.42	27.55	17.66	12.24	20.51	17.44%	23.33%	17.18%	5:24
TokenFlow	95.80	93.17	95.92	74.03	50.97	80.40	28.27	28.29	29.53	20.92	14.41	23.74	44.10%	68.97%	43.59%	5:24 (4.14)
Pix2Video	89.96	-	-	24.78	-	-	28.01	-	-	5.61	-	-	N/A	N/A	N/A	N/A
RAVE - w/o shuffle	93.98	89.90	92.49	71.78	47.26	76.58	28.78	29.49	29.71	20.66	13.94	22.76	N/A	N/A	N/A	N/A
RAVE	95.95	93.18	95.99	71.44	48.81	80.51	29.51	29.93	29.76	21.08	14.60	23.95	90.51%	82.82%	86.67%	4:28 (3:13)

longer videos, specifically at a frame length of 90. Moreover, when assessing comprehensive editing capability with Q_{edit}, we outperform other methods in all video lengths. Additionally, there is a significant increase in the CLIP-F score, particularly in longer videos (36 and 90 frames), when comparing RAVE w/o shuffling and RAVE. This emphasizes the importance of shuffling in preserving consistency.

User study While quantitative metrics offer a fair comparison among prior methods, assessing edit quality and temporal consistency remains challenging. Hence, we conduct a user study, posing questions (Q1: General Editing (GE), Q2: Temporal Consistency (TC), Q3: Textual Alignment (TA)) to 130 anonymous participants on a crowdsourcing platform, Prolific [5], for randomly selected 23 video-text pairs. Our survey results demonstrate superiority over previous methods in both aspects. TokenFlow, as anticipated due to its success in WarpSSIM score, secures the second-best results for temporal consistency.

Runtime We conduct a runtime comparison (shown in rightmost column in Table 1). Our results indicate that RAVE is ∼1 minute faster than the closest competitor, TokenFlow, on average for editing a single video of 90 frames, making it the preferred choice for fast video editing. Note that both our method and TokenFlow require preprocessing only once for a given video, regardless of the text prompt. In contrast, Text2Video-Zero uses random noises, and Rerender performs inversion only on the keyframes. Runtime without preprocessing durations are shared in Table 1, rightmost column in parenthesis.

Ablation study We conduct an ablation study by separately ablating 'shuffling', 'DDIM inversion', and ControlNet conditions (lineart, softedge and depth (RAVE)) in our framework, as illustrated in Fig. 7. Applying shuffling helps maintaining global style consistency (Fig. 7 (a)). Additionally, using DDIM inversion contributes to preserving the structure similar to the original image (Fig. 7 (b)). Furthermore, our approach proves to be adaptable to different controls, such as lineart (Fig. 7 (c)) and softedge (Fig. 7 (d)) compared to depth used in RAVE (Fig. 7 (e)). Even though there are style differences, these adjustments do not compromise the overall consistency.
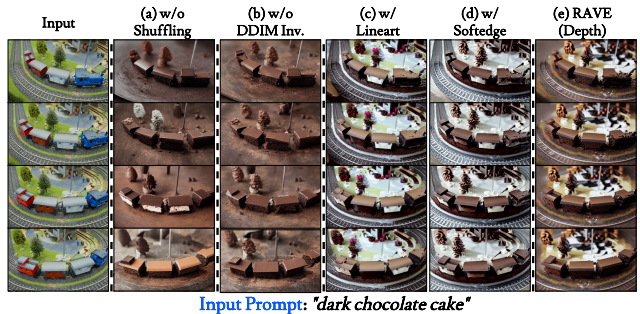

Input Prompt: *"dark chocolate cake"*

Figure 7. ***Ablation study.*** We ablate the following components: (a) shuffling, (b) DDIM inversion, and employ different conditions: (c) lineart and (d) softedge (e) depth.

Limitations While our quantitative evaluation clearly demonstrates our method's effectiveness in reducing flickering and achieving superior performance in comparison to existing methods (see Table 1), our method can occasionally leads to flickering when performing extreme shape editing or local attribute editing tasks (please see a detailed discussion in SM S.2) We attribute these issues primarily due to DDIM inversion's reliance on approximation, which accumulates errors with classifier guidance, and the compression phase in LDM. This leads to distortions and flickering in reconstructed images, especially in high-frequency areas.

5. Discussion and Conclusion

In this paper, we present RAVE, a lightweight, fast method for zero-shot text-guided video editing using T2I diffusion models, leveraging spatio-temporal interaction with reduced computational costs. Through extensive comparisons, RAVE outperforms previous baselines in terms of temporal consistency, speed, and alignment with textual editing, for video editing including longer ones. RAVE is adaptable to various pre-trained models (*e.g.* inpainting diffusion model, etc.), providing customizable video editing capabilities. We also highlight its potential for applications beyond video editing, such as consistent avatar generation or 3D texture editing, as part of our future work. By releasing our code and dataset for reproducibility, we encourage researchers to standardize evaluation.

References

[1] Civitai. https://civitai.com/. Accessed: 2023-11-16. 6

[2] Gridtrick. https://web.archive.org/web/20231025170948/https://semicolon.dev/midjourney/how-to-make-consistent-characters. Archived: 2023-10-25. 4

[3] Pexels. https://www.pexels.com/. Accessed: 2023-11-16. 1, 2

[4] Pixabay. https://pixabay.com/. Accessed: 2023-11-16. 1, 2

[5] Prolific. https://www.prolific.com/. Accessed: 2023-11-18. 8

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

[7] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2, 3

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2

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

[10] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 2, 3, 5, 7

[11] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 7, 1, 2

[12] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[14] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 3, 5, 7

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

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

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646. Curran Associates, Inc., 2022. 2, 3

[19] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023. 3

[20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[21] Ondrej Jamriska. Ebsynth: Fast example-based image synthesis and style transfer, 2018. 4

[22] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Transactions on Graphics (TOG)*, 38(4):1–11, 2019. 2

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

[24] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[25] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 2, 3, 5

[26] 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*, pages 14317–14326, 2023. 2, 3, 1

[27] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 2, 3

[28] 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. 3

[29] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 2

[30] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6038–6047, 2023. 3

[31] 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. 3

[32] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[33] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 1, 2

[34] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15932–15942, 2023. 2, 3, 5, 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

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

[37] 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. 2, 3, 4

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

[39] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[41] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 3

[42] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 4

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 7

[46] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 6

[47] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3

[48] 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. 3

[49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[50] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 3

[51] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023. 2, 3, 5, 7

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 4