electronics

MDPI

Article

Cross-Attention and Seamless Replacement of Latent Prompts for High-Definition Image-Driven Video Editing

Liangbing Zhao [1], Zicheng Zhang [2], Xuecheng Nie [3], Luoqi Liu [3,*] and Si Liu [4,*]

[1] School of Computer Science and Engineering, Beihang University, Beijing 100191, China; zhaolb@buaa.edu.cn
[2] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China; zhangzicheng19@mails.ucas.ac.cn
[3] Meitu Inc., 7th Floor, Block B/C, Yousheng Building, 28 Chengfu Road, Haidian District, Beijing 100083, China; nxc@meitu.com
[4] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
* Correspondence: llq5@meitu.com (L.L.); liusi@buaa.edu.cn (S.L.)

Abstract: Recently, text-driven video editing has received increasing attention due to the surprising success of the text-to-image model in improving video quality. However, video editing based on the text prompt is facing huge challenges in achieving precise and controllable editing. Herein, we propose Latent prompt Image-driven Video Editing (LIVE) with a precise and controllable video editing function. The important innovation of LIVE is to utilize the latent codes from reference images as latent prompts to rapidly enrich visual details. The novel latent prompt mechanism endows two powerful capabilities for LIVE: one is a comprehensively interactive ability between video frame and latent prompt in the spatial and temporal dimensions, achieved by revisiting and enhancing cross-attention, and the other is the efficient expression ability of training continuous input videos and images within the diffusion space by fine-tuning various components such as latent prompts, textual embeddings, and LDM parameters. Therefore, LIVE can efficiently generate various edited videos with visual consistency by seamlessly replacing the objects in each frame with user-specified targets. The high-definition experimental results from real-world videos not only confirmed the effectiveness of LIVE but also demonstrated important potential application prospects of LIVE in image-driven video editing.

Keywords: video editing; diffusion model; attention control

check for updates

Citation: Zhao, L.; Zhang, Z.; Nie, X.; Liu, L.; Liu, S. Cross-Attention and Seamless Replacement of Latent Prompts for High-Definition Image-Driven Video Editing. *Electronics* **2024**, *13*, 7. https://doi.org/10.3390/electronics13010007

Academic Editor: Stefanos Kollias

Received: 10 November 2023
Revised: 9 December 2023
Accepted: 15 December 2023
Published: 19 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, video generation has gone through the initial stage of generating a video from noise [1–4] and the rapid development stage of video editing [5–8] based on the latest research achievements of AI-Generated Content (AIGC) and various rapid generative models, such as diffusion models [9] and GANs [10]. Much important progress has been made in video editing research. It has been extensively reported that the attention control between text and images manipulates and substitutes objects and styles within video sequences [11,12]. The important role of null-text inversion in maintaining video background consistency has been discovered [13]. However, the existing video editing tasks are limited to text-based modifications, which cannot meet the customized needs expected by users. Especially, up to now, the possibility of directly incorporating input images into the video editing process has not been explored and reported.

To address these issues in the video editing research above, our paper presents an image-driven video editing model that allows users to change specific parts of a video according to the user's real-time input images. The proposed solution can provide a new method for more precise personalized video editing. To achieve image-driven video editing, we have completed a series of challenging tasks, involving preserving the consistency of across-frame videos, adhering to the sequence of motion videos, and replacing an object

within input images with a corresponding video component. Unlike traditional video editing, our image-driven video editing requires us to simultaneously input and edit both images and text.

To achieve the image-driven video editing task, we have designed an image-driven video editing framework based on the text-to-video (T2V) model. Compared to the Tune-A-Video (TAV) [14], our T2V model can generate more abundant motion information, increase the continuity between frames in a video, and thus achieve high-definition video editing tasks under various image conditions. However, the newly built T2V model does not take into account the information contained in input images, which can be solved by two steps as follows. Firstly, it needs to fine-tune the original video based on the learning motion and background. Secondly, it needs to incorporate the input images into our T2V model and train them to achieve the alignment of text and image. In addition, it also requires an attention control method during the inference process to maintain background consistency and generate final edit results.

The key contributions of this paper are as follows:

(1) We introduce a novel image-driven video editing task in order to generate customized videos based on input images;
(2) We propose a LIVE framework for the customized video editing task and delicately design a training and reasoning pipeline to successfully overcome a series of challenging tasks that occur during the image-driven video editing process;
(3) Extensive experiments demonstrate the effectiveness of our LIVE and obtain promising results in multiple aspects such as image object swap, video consistency, and motion maintenance (Figure 1).

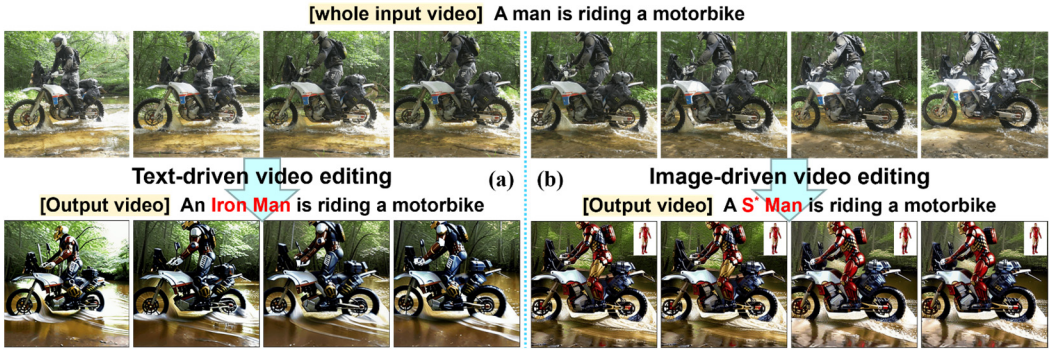


Figure 1. The difference between text-driven video editing and image-driven video editing. (**a**) Text-driven video editing uses text prompts to modify input videos for an iron man and can only obtain low-definition output videos. (**b**) Image-driven video editing uses an image prompt to delicately modify input videos for an S* man and can produce high-definition output videos.

Our T2V basic model is implemented by extending 2D stable diffusion. In the first step, we not only modified the T2V framework but also trained the motion information of input videos with a learnable latent prompt. In the second step, we not only aligned text and image but also established a mapping between text and image with the cost-effective leverage of textual inversion [15]. We also designed a pipeline to accelerate the training process and improve the training effectiveness. In order to obtain high-definition edited videos, the final inference stage not only needs to replace the latent prompt and text with the relevant training outcomes but also needs to utilize custom attention control.

2. Background

2.1. Image Generation and Editing

In recent years, much important progress has been made in image generation research. Some well-known generative models, such as GANs [10], have demonstrated powerful abilities to generate various realistic images [16,17]. However, the denoising diffusion

probabilistic models (DDPMs) [9] recently received increasing attention because of the high-quality output ability for large-scale datasets [18]. The DALLE-2 [19] could further enhance the text–image alignments by utilizing the CLIP feature space [20]. Imagen [21] employed cascaded diffusion models [22] for high-fidelity image generation. The latent diffusion models (LDMs) [23] were introduced to manage the generation process in the latent space and boost training efficiency.

On the other hand, more and more exciting image editing works have also been reported. The text-to-image (T2I) diffusion models had first taken away the dominant position determined by previous state-of-the-art models for text-driven image editing [24,25]. Recently, some works based on a pre-trained T2I diffusion model have achieved remarkably enhanced performance in text-driven image editing [26–29]. The prompt-to-prompt (P2P) accomplished the precise control of the spatial layout in various edited images by directly manipulating the text–image cross-attention mappings during the image generation process [11]. InstructPix2Pix [30] and Paint-by-Example [29] enabled us to characterize image editing according to the user-provided instructions. Textual inversion [15], DreamBooth [31], and XTI [32] had the preliminary ability to learn special tokens for personalized concepts and generate the corresponding images. Null-text inversion [13] carried out null-text optimization for real image editing.

2.2. Video Generation and Editing

Unlike image generation, video generation has been facing huge challenges because of the higher-dimensional complexity and the scarcity of high-quality datasets. The Video Diffusion Models [1] generated temporally coherent videos with a 3D U-Net by decomposing space–time attention [33]. Imagen Video [3] utilized a cascaded Video Diffusion Model to achieve high-resolution video generation. However, these methods require paired text–video datasets; thus, the training cost is very high. Recently, the TAV [14] achieved one-shot tuning video generation by transforming a T2I model into a T2V model and fine-tuning input videos.

The T2I diffusion models had strong image editing ability but no exploration of video editing. Text2Live [34] demonstrated compelling video editing results by combining layered neural representations and text guidance [35]. Dreamix [5] performed image-to-video and video-to-video editing and had the simple ability to change motion by employing a pre-trained Imagen Video backbone [3]. Gen-1 [36] accomplishes some stylization and customization tasks by simultaneously training images and videos. However, these methods require costly training. To address the issue, various fine-tuning methods have been proposed. Based on the P2P image editing method [11], Video-P2P was developed to modify the attention map corresponding to the text prompt and maintain the consistency of other parts of the attention heatmap using local blending [7]. It was also proposed to decrease the control over the guiding attention and enhance the preservation of the unedited area. Edit-A-Video [37] also used the attention replacement mechanism in P2P and null-text inversion [13] to edit videos and maintain coherence. Pix2Video [6] introduced self-attention feature injection, which could maintain the edited videos' appearance coherence. It further developed latent guidance using the l2 loss as a guidance of the ground truth frame between the predicted current frame and the previous frame, remarkably improving the video temporal consistency. ControlVideo [38] extended the structural capabilities of ControlNet to the video domain, enabling more sophisticated video editing functionalities. StableVideo [39] designed an interframe propagation mechanism to generate coherent geometry objects and used NLA to achieve temporal consistency. CCEdit [40] harnessed structure control with a pre-trained ControlNet and achieved appearance control with extended text prompts, personalized T2I models, and edited center frames.

These achievements are the most representative of progress in text-driven video editing; however, their video editing is still at the level of modifying the text prompts. All the research methods cannot be applied to the image-driven video editing task. LIVE is

proposed in this paper for the image-driven video editing task by modifying the TAV and the attention control.

3. Methods

In Section 3.1, we first briefly introduce the DDPMs [9]. In Section 3.2, we comprehensively introduce LIVE framework, which has been designed to achieve the image-driven video editing task through a latent prompt. In Section 3.3, we, in detail, explain how LIVE generates various edited videos. In Section 3.4, we explain how LIVE controls the attention to achieve high-definition edited videos.

3.1. Preliminary

The DDPM [9] is recently a popular generative model consisting of a forward noising process $q(x_t|x_{t-1})$ and a backward denoising process $p_\theta(x_{t-1}|x_t)$, where t is the timestep x_t and x_{t-1} are the noisy input data at timestep t and $t-1$, respectively, and θ is the neural network to optimize. The goal is to first add noise to images and converge to Gaussian noise $z \sim N(0,1)$ step by step and then to denoise images to obtain the restored images. This can be expressed as:

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

where ε_t is the random noise, x_0 is the original image, and β_t is a hyperparameter of the forward process, $\alpha_t = 1-\beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. The way to choose $\overline{\alpha}_t$ is called the noise schedule. We chose a cosine noise schedule as mentioned by Nichol and Dhariwal [18]. When x_0 is known, $q(x_t|x_{t-1})$ in the Equation (1) can be expressed as:

$$q(x_t|x_t,\ x_0) := N\left(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_t\right), \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t\right)$$

During the denoising process, it is necessary to use the equation $q(x_{t-1}|x_t)$ to handle the reverse process. However, we need a deep neural network to learn the denoising distribution because the equation $q(x_{t-1}|x_t)$ cannot be solved. The denoising distribution can be defined as:

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t,t), \sum\nolimits_\theta(x_t,t)) \tag{2}$$

where $\mu_\theta(x_t,t)$ is expressed as $\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_t\right)$. For the $\sum_\theta(x_t,t)$ part, Nichol and Dhariwal [18] found that it had upper and lower limits; therefore, in order to simplify the training objective, it can be set as a constant within a specific region as follows:

$$L_t = \left\| \epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon_t, t\right)\right\|^2 \tag{3}$$

The DDPM's reverse process is a stochastic process and unstable, thus easily leading to different reconstructed results from the same inputs. Such behavior could have disastrous consequences for video editing tasks since video editing requires only partial changes within a specific region and must not destroy the original appearance of other parts of the edited image. Fortunately, the introduction of the DDIM [41] had successfully resolved this problem. During the inference process, the DDIM set $\sum_\theta(x_t,t)$ to be 0; as a result, the inference process is transformed from a stochastic process to a deterministic one. Therefore, a denoising process can be accomplished through fewer steps under the condition of ensuring consistency between the denoised results and the input images, except for a slight decrease in image quality. This denoising approach, called the DDIM inversion, has been integrated into our LIVE method.

3.2. LIVE Framework Overview

LIVE framework is depicted in Figure 2a,b. It can be seen that LIVE framework starts with the T2I model, as shown in Figure 2a, which is subsequently extended into a T2V model. In terms of technical implementation, we chose the configuration of TAV [14] because TAV is one of the few widely recognized and available T2V models. In the original Latent Diffusion Model (LDM), the attention mechanism comprises spatial self-attention and cross-attention components. To extend this to video, we first inflate 2D convolution kernels to 3D, thereby incorporating the temporal dimension. Second, following the Video Diffusion Model (VDM) baseline, we incorporate factorized space–time attention, adding temporal attention subsequent to the cross-attention mechanism to learn sequential information effectively. Temporal attention, as utilized in our framework, follows the factorized space–time attention mechanism as established in the Video Diffusion Models (VDMs). It operates fundamentally in the same way as the original spatial self-attention, with the key difference being the dimension it operates over. While spatial self-attention computes attention across the spatial dimension, temporal attention is calculated over the time dimension, enabling the model to capture and utilize temporal dynamics for video editing tasks.

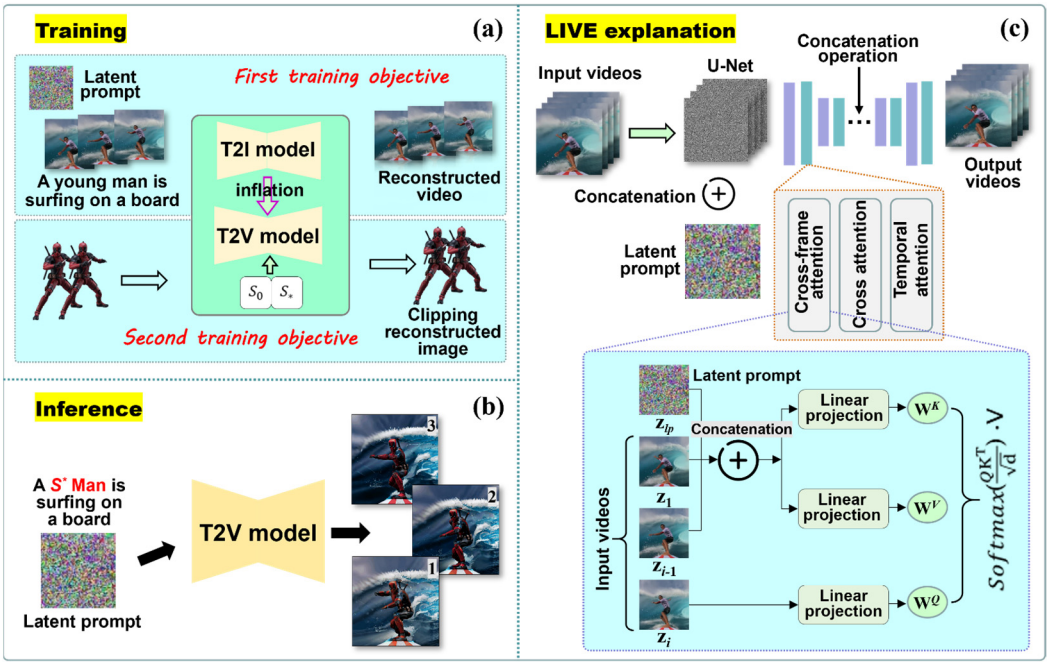


Figure 2. Illustration of the proposed LIVE framework. (**a**) Overview of training: The T2I models are first inflated to the T2V models, and then the T2V models are trained with two objectives. The first objective training is that using the latent prompt fine-tunes text–video pairs to obtain the trained T2V models. The second training objective is that the input images are aligned with specific text that is embedded to achieve personalized video editing abilities. (**b**) Overview of inference: The T2V models trained by two objectives use modified text prompts to generate a series of edited videos. (**c**) An extensive explanation of LIVE: The core idea of our LIVE is to use the input videos as fine-tuning objects. We need to first obtain a noisy latent during the diffusion forward process, and then the noisy latent is concatenated with the latent prompt in frame dimension, and lastly, the concatenated boy is seen as an input of U-Net. We still need to update the attention block with the diffusion reconstruction loss. In the cross-frame attention blocks, all the features of the latent prompt z_{lp}, the first frame z_1, and the previous frame z_{i-1} are linearly projected to the K key (W^K) and the V value (W^V), while the features of the current frame z_i are projected to the Q query (W^Q). All the outputs are sent to cross-frame attention blocks to update z_i.

However, this T2V model alone is insufficient for accomplishing image-driven video editing tasks because it cannot generate coherent videos and also cannot establish mapping with the input images. To realize image-driven video editing, we trained the T2V model with two objectives as follows. The first objective of training is to acquire video reconstruction ability. At this stage, the input information includes the videos to be fine-tuned and the corresponding text prompt. These videos are first encoded into latent prompts by the Variational AutoEncoder (VAE). Then, the latent prompts are given noise and concatenated with a learnable tensor ($1 \times 64 \times 64 \times 4$) along the frame dimension. Except for the frame dimension, the shape of the latent prompt is the same as the latent video. We will obtain a reconstructed video by inputting the concatenated tensor into U-Net. The main goals of the second objective of training are to incorporate the input image information into the T2V model and try to associate all the input images with a specific space in the text embeddings. The aim of these operations is to ensure the two objectives' training costs are at the same level and create a favorable condition in which the two training objectives are simultaneously converged to their optimal state by an alternating training method. The final training objective is:

$$L_t^i = \begin{cases} L_t^{unet}, \ when \ i = 2n \\ L_t^{text}, \ when \ i = 2n + 1' \end{cases} \tag{4}$$

where i denotes the training iteration and $n = 1, 2, 3, \ldots$ L_t^{unet} is the diffusion loss of optimizing U-Net, L_t^{text} is the diffusion loss of optimizing the text embeddings, and λ is the loss coefficient of balancing the effect of the two objectives training. Experimentally, λ is usually set to 1. The specific numerical forms of L_t^{unet} and L_t^{text} are the same, which can be expressed in Equation (3).

After having accomplished the two training objectives above, the T2V model will be used to carry out image-driven video editing through a straightforward process. As shown in Figure 2b, we first need to substitute the object name of the text prompt in the original video with the rare string S_* and then set the S_* as an input of the inference. Meanwhile, at the inference stage, we utilize the attention control mechanism in order to obtain the final edited videos.

The innovation of LIVE framework is to introduce the latent prompt into the T2V model. This is where the T2V model is completely different from the TAV model [14]. We discovered that the accurate object replacement and the seamless motion flow of the replaced object determine whether the image-driven video editing task can be successfully completed. Therefore, the motion information for learning the T2V model must be adequate to support the object's movements. The aim of introducing latent prompts is to add learnable variables to enhance the motion perception of the T2V models. Therefore, the latent prompts mechanism provides a good solution for the problem of insufficient motion information in the T2V model.

3.3. Latent Prompt Video Editing

Latent prompt. The frame manipulation of videos is a highly challenging but currently unsolved research topic. To explore the possible approach of manipulating the frames in a video, we introduce the latent prompt that can be fine-tuned on spatial and temporal dimensions. This spatial–temporal interactive fine-tuning is an important guarantee for achieving continuous and seamless video editing.

Given the video latent vector $z \in R^{f \times h \times w \times c}$, the latent prompt z_{lp} for one frame in the same latent space as z can be defined as $z_{lp} \in R^{1 \times h \times w \times c}$. Based on these conditions, we can initialize the latent prompt with the first frame of the input video. The spatial–temporal interactive fine-tuning starts with the latent prompt performing the cross-attention on all the frames in each video, enabling them to learn motion information effectively. Obviously, this spatial–temporal interactive fine-tuning can make z_{lp} grasp complex spatial details and various relationships between different frames in a video. According to the

spatial and temporal information from various frames, the latent prompt can provide more comprehensive visual information for the T2V models. This is very important for the T2V models to more precisely manipulate the video's frames. The temporal attention of the latent prompt can play a vital role in learning and encoding the motion's dynamic information of the video sequence. With the help of cross-attention mechanisms, the latent prompt can fully understand the temporal dynamic characteristics of a video.

Training. To achieve the first training objective, the latent prompt has always been a tensor in the latent space that can be optimized, which is similar to the text prompt mechanism used in the text encoder. As shown in Figure 2c, the latent prompt is concatenated with the latent vector of input videos along the frame dimension condition, which is input into U-Net for the second training objective. We can modify the attention mechanism by incorporating the latent prompt z_{lp}, the first frame z_1, and the previous frame z_{i-1} as a new attention of input video, shown at the bottom of Figure 2c. Our attention mechanism can be expressed as:

$$Q = W^Q z_i, K = W^K \left[z_{lp}, z_1, z_{i-1}\right], V = W^V \left[z_{lp}, z_1, z_{i-1}\right] \tag{5}$$

where the $[\cdot]$ in Figure 2c denotes the concatenation operation and W^Q, W^K, and W^V are projection matrices. Our attention mechanism can ensure the latent prompt to maintain the cross-frame attention with each frame in all sequences. Therefore, our latent prompts can be enriched with massive video motion information, which can enhance the naturalness and smoothness of the motion videos generated by the T2V model.

To achieve the second training objective, we create an alignment between the rare strings, such as "$\langle aka* \rangle$", and images by providing rare strings aligned with the embedded layer value of the initial token for initialization using the textual inversion [15]. We use the original image search from the Internet as the input video without any segmentation. To introduce the latent prompt, it is necessary to incorporate a latent prompt into the input. As shown in Figure 2a, our input video clip is a combination of the input image and its copy. Herein, the input image represents the latent prompt, and the copy represents the standard input for textual inversion. We discover that the temporal attention in U-Net has a negative impact on subsequent learning. Consequently, it is necessary to not use temporal attention during the second training objective process to achieve optimal training results.

To achieve a balance between pursuing high-definition video and saving training time, we have adopted a strategy of synchronously updating the latent prompt and U-Net architecture. However, during the second training objective process, we freeze the latent prompt and replace it with the latent vectors of the input image. We discover that the cross-attention mechanism, which takes both text and video as the inputs, and outputs the latent prompt for training, cannot obtain satisfactory results. This failure may originate from the limited training steps. Therefore, we replace the previous complex mechanism with a learnable variable to create a fine-tuned model, leading to better results and also improving training efficiency.

Inference. After having completed the two training objectives, we can directly obtain the edited video from the rare strings used during training. However, we need to analyze the processing process of the latent prompts. During the first objective training process, we trained the latent prompt using a trainable vector and we replaced the latent prompt position with the latent image during the second objective training process. The inference of the diffusion model generally adopts the guidance method unrelated to classifiers [42]. For the conditional part, we set the input to the trained latent prompt and noise in order to enhance the motion information required for the input images. When it comes to the unconditional part, we set the input as the latent image and noise so that the generated videos cannot only include the image information bound by the text prompt through the second part of LIVE framework but also supplement the input with cross-frame attention. Therefore, these operations can generate high-definition edited videos with content richness and visual consistency.

3.4. Attention Control

After having trained the T2V model with LIVE framework, we can perform object swapping. Attention control has been widely applied because it does not require time-consuming training editing. Attention control can guide the T2V model to generate background-preserved images by swapping the maps of self-attention and cross-attention layers under the condition that all the specific information of the original image is completely preserved. We can have successfully accomplished the object replacement in the no objects area under the condition of maintaining the consistency between the original video and its edited version. After obtaining the mask M_t of the origin prompt attention and mask M_t^* of the edited prompt attention at timestep t with the settings of P2P, we can adopt the word swap edit function $Edit\left(M_t, M_t^*, t\right)$, which can be expressed as:

$$Edit\left(M_t, M_t^*, t\right) := \begin{cases} M_t^* \; if \; t < \tau \\ M_t \; otherwise \end{cases} \tag{6}$$

where τ denotes the starting of attention injection. Based on the observation that the early steps of diffusion models can produce an important contribution to the overall distribution of the final edited videos, we inject a mask M_t^* of the edited attention to guide the video generation. Since the original word swap can only be applied to those text prompts with the same length, we modify it to specify which part of the text prompt needs replacement. As a result, we created a flexible plan for Word Swap. It should be pointed out that we only need to replace the original word swap with verbs representing motions and nouns representing background in the sentence. We did not choose to replace the object because we hope the edited video contains accurate objects from the input images. The reason why we replace verbs representing motions is that we do not want the object information in the origin attention map to interfere with our editing videos. The reason why we replace nouns representing backgrounds is that we need to ensure that the generated videos can match the new background. However, we discover that only replacing the background is not sufficient because the background's attention map is not as accurate as the object's attention map. VideoP2P [7] and other text-driven video editing methods [37] try to utilize null-text inversion [13] to solve this problem, but these approaches will greatly increase the model training cost. Therefore, in order to achieve promising high-definition image-driven videos, we choose to first use the local blend method in P2P [11] to compulsively replace the background and then utilize Gaussian blur to adjust the transition between the objects and the backgrounds.

4. Results

Implementation. We have implemented our approach based on the stable diffusion v1-41. For our video samples, we adopted a frame sampling approach where we extracted one frame every five frames, resulting in a total of eight sampled frames per video. Similar to the TAV [14], we have fine-tuned the attention modules on an eight-frame and 512×512 video in U-Net architecture for 500–600 steps to optimize U-Net. Our learning rate is 3×10^{-5}. For the textual inversion component, we use only a single input image and set the placeholder marker to "<@#$>" and have trained it using a learning rate of 5×10^{-3} for 500–600 steps. For the latent prompt, we have initialized it with the first frame of the original video during the first training objective process; however, it is fixed at the latent state of the input image during the second training objective process. During the inference process, we first save DDIM inversion values [41] and set the classifier-free guidance to 12.5 [42]. Then, we completed 50 steps of inferring operation using the DDIM sampler. In terms of attention control, we set the cross-attention replacing ratio to 0.6 and the self-attention replacing ratio to 0.3. All the experiments have been conducted on an RTX 3090ti GPU. Training 500 steps takes 15 min. The inferring time to obtain the DDIM inversion values takes 10 min.

Baselines. Since our proposed task has not been explored before, we choose several concurrent text-driven video editing methods as a baseline for comparison. These baseline methods include (1) TAV + DDIM [14]. This method is used to generate videos from text prompts by fine-tuning the pre-trained T2I diffusion models using an efficient and one-time tuning approach. (2) Video-P2P [7]: Based on the TAV models, this method is used to control the video content through text editing using attention control and null-text inversion. (3) Vid2Vid-zero [43]: This approach is used to achieve zero-shot video reconstruction through abundant spatial–temporal attention modules and video editing by simultaneously using the attention control and the null-text inversion. (4) Paint-by-Example [29]: This method is an image-based image inpainting approach. It utilizes a mask-conditioned diffusion model to achieve seamless inpainting. In our comparative analysis, we adapted Paint-by-Example to video editing by applying the inpainting process to each frame of a video sequence. All the experiments have been conducted with their official codes and configurations.

Dataset. We chose the conventional videos used in Video-P2P [7] and Vid2Vid-zero [43] as the evaluation dataset. We first used their default text prompts to edit videos and then fairly compared the results.

4.1. Main Results

Qualitative results. Figure 3 shows a set of examples achieved by our method. For each example, we show the second, fourth, sixth, and eighth frames of the input video and the edited output videos, as well as the corresponding text prompt and input images. It can be seen that LIVE is capable of handling these videos with prominent foreground objects and multiple foreground objects. We can accurately replace the objects with the input image. After having completed the replacement operation, the motion features of the edited videos are very natural, which are the same as those in the original video. Furthermore, our method can handle a wide range of object replacements. For example, as demonstrated in the second row, a motorcycle being ridden by a person can be replaced by a horse, or as shown in the fourth row, a car can be replaced by a train. These examples demonstrate the effectiveness of our LIVE method.

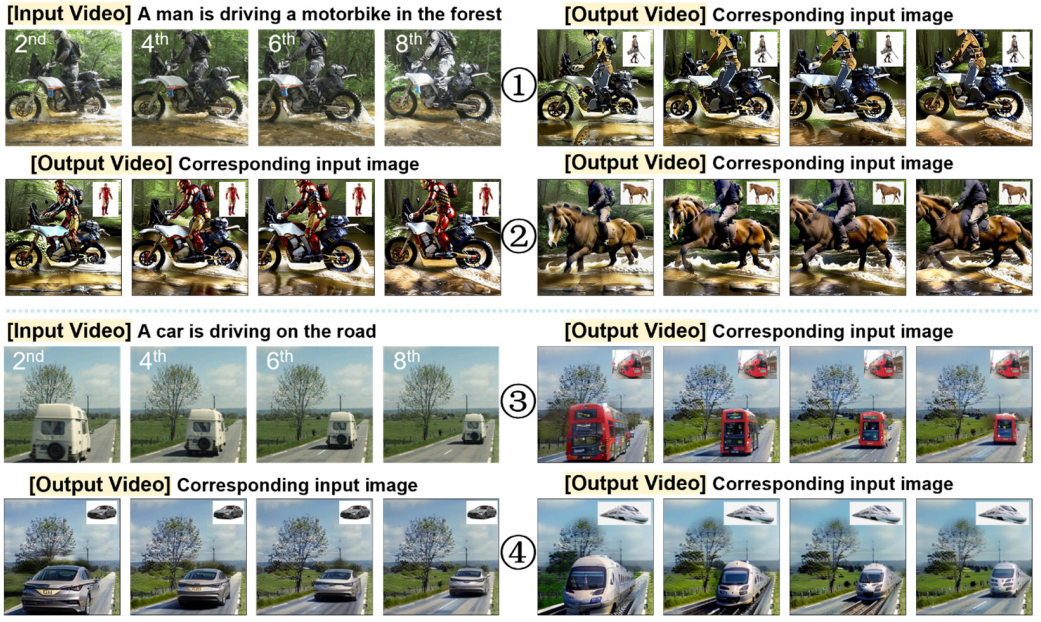


Figure 3. LIVE results from various input videos, images, and prompts. LIVE can accurately replace the designated object in the input image with the corresponding object in the video under the condition of maintaining the visual continuity of the edited video.

Comparisons. Figure 4 shows the four frames of the input video and output videos edited by LIVE, Vid2Vid-zero, Video-P2P, TAV + DDIM*, and Video-P2P*, and Paint-by-Example, where the $*$ means that the method has been implemented with textual inversion. Vid2Vid-zero is a training-free method and is thus not suitable for adding the textual inversion. All other baseline methods are designed for text-driven video editing. By comparing with LIVE editing videos, all other baseline methods, except for Vid2Vid-zero, can achieve textual inversion and generate the input objects. However, it can be seen that their image-driven video editing performance is not very satisfactory. Vid2Vid-zero can generate some general videos that basically reflect the overall picture but have poor motion. Video-P2P without textual inversion can generate much better videos, in which the background is well-preserved and impressive; however, the motion and pose of the object are blurry. TAV + DDIM can effectively replace the object in the input image into the video, thus obtaining the best edited videos. However, all the edited videos obtained by three baseline methods still maintain the original content of the video in the background. The image-based inpainting method Paint-by-Example is clearly weaker than ours in aspects such as temporal coherence and editing effects due to a lack of temporal information. It is worth noting that our LIVE approach can successfully achieve both object replacement and background preservation. The edited videos generated by the LIVE approach exhibit smoother and more continuous motion and thus have better overall video coherence.

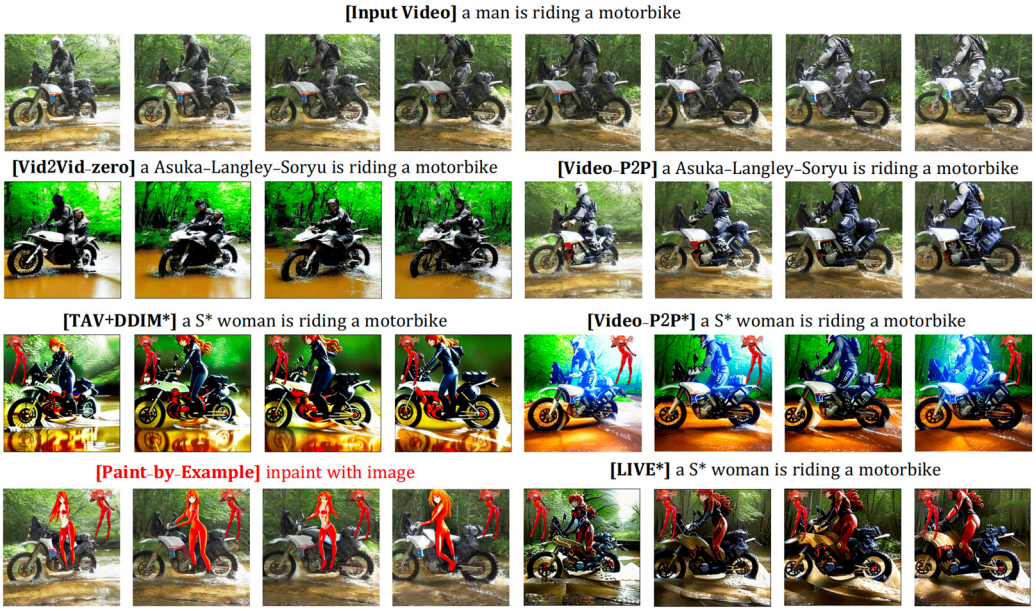


Figure 4. Comparison of different edited videos by LIVE and the baseline methods. The $*$ means that the method has been implemented with textual inversion. Only LIVE achieves both accurate object replacement and video consistency.

Quantitative results. (1) CLIP score: To evaluate the text–video alignment, we choose the setting of Imagen Video [3] to calculate the CLIP score [44]. We first calculated at the frame level the CLIP scores of four methods and then calculated the average value of the calculated CLIP scores for each method and took the average value as the CLIP score of the method. As shown in Figure 5a, the CLIP score of LIVE is the highest of all the methods, demonstrating that our method possesses better text–video alignment. (2) User study: We further evaluate LIVE and the baseline methods based on a user study. The questionnaire includes five videos, each of which has been edited using three kinds of baseline methods and our LIVE. For each video, there are two questions. The first is which method best preserves the input image after having completed video editing? The second is which method has the best-preserved background after having completed video editing? The two questions correspond to the two challenges of our proposed image-driven video editing

task, which are how to achieve accurate object replacement and video coherence. Since Paint-by-Example is an image-based inpainting method, comparing it with other methods on quantitative metrics, like background preservation and similarity, to the input image is obviously unfair. Therefore, we did not present the quantitative results of this method. As shown in Figure 5b,c, the majority of the participants agree that our LIVE method can more accurately replace objects than any other baseline methods. For background preservation, our LIVE method beat other baseline methods, except for the Video-P2P method, because the Video-P2P method utilizes null-text inversion and delicately devises an attention control mechanism to achieve better preservation results. Therefore, the user study confirms that our LIVE method performs much better than other baseline methods in object replacement.

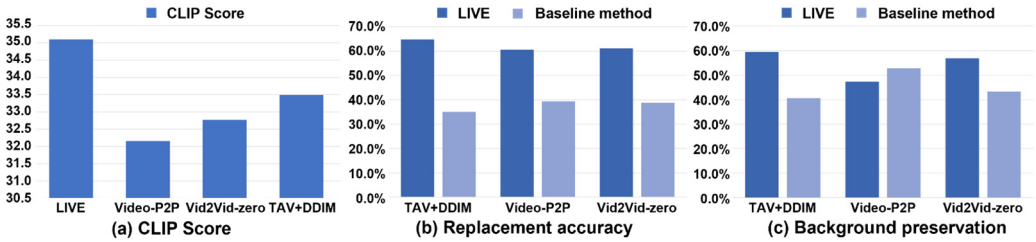


Figure 5. Quantitative results obtained through the user study. The results show that our LIVE framework can obtain better performance in object replacement accuracy, background preservation, and clip score than other baseline methods.

4.2. Ablation Results

Latent prompt. Since our LIVE focuses on the latent prompt, we have first examined the influence of the latent prompt on the overall framework. As shown in Figure 6, it can be seen that under the case without a latent prompt, the texture and shape of the object cannot be well preserved, and there are some strange motions appearing in the output of edited videos, as shown in the red-line circles. After having added the latent prompt, precise object replacement has been achieved, and the motion becomes more continuous. Especially important is that we have not observed any strange motions in all the output edited videos with the latent prompt.

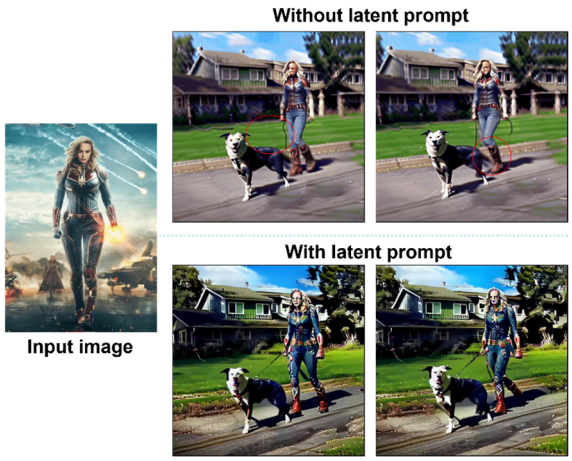


Figure 6. Ablations on latent prompts.

Training of temporal attention. During the second objective training process, our goal is to bind the text to the input image. Theoretically, this training should have been carried out on a 2D model without the time dimension. However, since our LIVE model has been extended to the 3D model, temporal attention must be involved during the whole training process. More intuitively speaking, temporal attention should not be trained

during the second objective training process; otherwise, both the temporal attention and the binding quality between the text and the input images will be significantly affected. As shown in Figure 7, when temporal attention is trained during the second objective training process, the operation will not only make image learning not satisfactory but also lead to the appearance of artifacts. We discovered that freezing the temporal attention can significantly improve the overall results during the second objective training process.

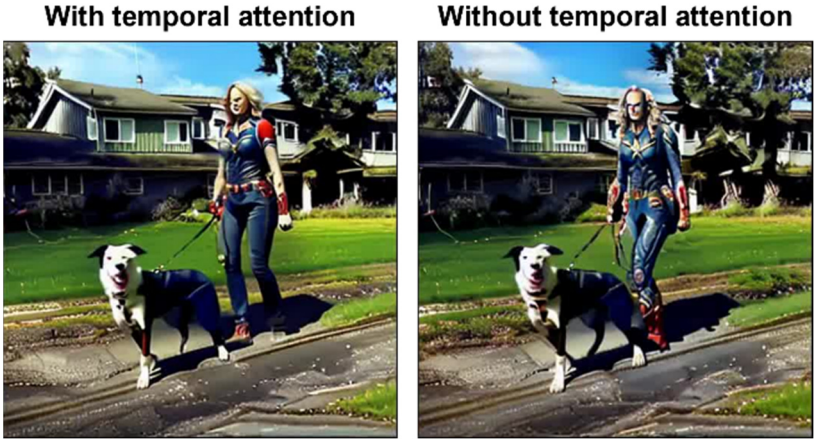


Figure 7. Ablations on temporal attention.

Attention control. As shown in Figure 8, we can clearly observe the difference in background preservation with and without attention control. Without attention control, the background becomes noticeably darker compared to the original video, and the water splashes under the motorcycle tires are not adequately represented. However, these issues can be greatly alleviated using attention control, indicating the critical role of attention control, especially in preserving the original background while editing specific objects in the video.



Figure 8. Ablations on attention control.

5. Conclusions

In summary, a novel image-driven video editing (LIVE) framework is proposed in this paper. Based on LIVE framework, the image-driven video editing task has been successfully achieved in various video editing scenarios by replacing objects in various edited videos with input images. The important innovation of LIVE is to utilize the latent codes from reference images as latent prompts to rapidly enrich visual details. The novel latent prompt mechanism endows LIVE with two powerful capabilities: one is the comprehensively interactive ability between the video frame and latent prompt in the spatial and temporal dimensions, which is achieved by revisiting and enhancing cross-attention, and the other is the efficient expression ability of training continuous input videos and images within the diffusion space by fine-tuning various components such as latent prompts, textual embeddings, and LDM parameters. LIVE can efficiently generate various edited videos with visual consistency by seamlessly replacing the objects in each frame with user-specified targets. The high-definition editing videos from the demonstration experiment confirm that LIVE can precisely control the video editing process and achieve impressive performance on

real-world images. This work provides a flexible way for the modification and generation of the edited videos and is expected to serve as a solid baseline for image-driven video editing research.

The major limitation of the proposed LIVE framework is its constrained ability to handle edits that involve substantial shape changes or cross-species transformations. This limitation arises due to the reliance on attention masks for object replacement. When the shape variation is too pronounced, the attention mask may not fully cover the object, leading to unsuccessful edits. This insight has led us to consider future developments that might employ methods like gradient-guided editing (e.g., DDS), which could overcome the limitations imposed by attention masks. Additionally, exploring pure video generation models might further improve editing capabilities, especially for extensive modifications.

Author Contributions: Methodology, L.Z. and Z.Z.; Validation, L.Z., Z.Z. and X.N.; Writing-original draft, L.Z. and Z.Z.; Writing-review & editing, X.N., L.L. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China: 62122010.

Data Availability Statement: Data is open-sourced on GitHub.

Conflicts of Interest: Author Xuecheng Nie and Luoqi Liu was employed by the company Meitu Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video diffusion models. *arXiv* **2022**, arXiv:2204.03458.
2. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-a-video: Text-to-video generation without text-video data. *arXiv* **2022**, arXiv:2209.14792.
3. Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D.P.; Poole, B.; Norouzi, M.; Fleet, D.J.; et al. Imagen video: High-definition video generation with diffusion models. *arXiv* **2022**, arXiv:2210.02303.
4. Harvey, W.; Naderiparizi, S.; Masrani, V.; Weilbach, C.; Wood, F. Flexible diffusion modeling of long videos. *arXiv* **2022**, arXiv:2205.11495.
5. Molad, E.; Horwitz, E.; Valevski, D.; Rav, A.A.; Matias, Y.; Pritch, Y.; Leviathan, Y.; Yedid, H. Dreamix: Video diffusion models are general video editors. *arXiv* **2023**, arXiv:2302.01329.
6. Ceylan, D.; Huang, C.-H.P.; Mitra, N.J. Pix2video: Video editing using image diffusion. *arXiv* **2023**, arXiv:2303.12688.
7. Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; Jia, J. Video-P2P: Video editing with cross-attention control. *arXiv* **2023**, arXiv:2303.04761.
8. Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv* **2023**, arXiv:2303.09535.
9. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
10. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
11. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv* **2022**, arXiv:2208.01626.
12. Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.* **2023**, *42*, 1–10. [CrossRef]
13. Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. *arXiv* **2022**, arXiv:2211.09794.
14. Wu, J.Z.; Ge, Y.; Wang, X.; Lei, W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; Shou, M.Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv* **2022**, arXiv:2212.11565.
15. Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A.H.; Chechik, G.; Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv* **2022**, arXiv:2208.01618.
16. Brock, A.; Donahue, J.; Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv* **2018**, arXiv:1809.11096.
17. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
18. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
19. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text- conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
20. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, L.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual only, 18–24 July 2021; pp. 8748–8763.

21. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Lopes, R.G.; Ayan, B.K.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
22. Ho, J.; Saharia, C.; Chan, W.; Fleet, D.J.; Norouzi, M.; Salimans, T. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.* **2022**, *23*, 1–33.
23. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High- resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 10684–10695.
24. Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; Cohen-Or, D. Encoding in style: A stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2287–2296.
25. Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; Cohen-Or, D. Designing an encoder for stylegan image manipulation. *ACM Transac. Gr. (TOG)* **2021**, *40*, 1–14.
26. Couairon, G.; Verbeek, J.; Schwenk, H.; Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv* **2022**, arXiv:2210.11427.
27. Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; Irani, M. Imagic: Text-Based Real Image Editing with Diffusion Models. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 6007–6017.
28. Tumanyan, N.; Geyer, M.; Bagon, S.; Dekel, T. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 1921–1930.
29. Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; Wen, F. Paint by Example: Exemplar-based Image Editing with Diffusion Models. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18381–18391.
30. Brooks, T.; Holynski, A.; Efros, A.A. InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18392–18402.
31. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 22500–22510.
32. Voynov, A.; Chu, Q.; Cohen-Or, D.; Aberman, K. p+: Extended textual conditioning in text-to-image generation. *arXiv* **2023**, arXiv:2303.09522.
33. Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv* **2019**, arXiv:1912.12180.
34. Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; Dekel, T. Text2LIVE: Text-Driven Layered Image and Video Editing. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 707–723.
35. Lu, E.; Cole, F.; Dekel, T.; Xie, W.; Zisserman, A.; Salesin, D.; Freeman, W.T.; Rubinstein, M. Layered neural rendering for retiming people in video. *arXiv* **2020**, arXiv:2009.07833. [CrossRef]
36. Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; Germanidis, A. Structure and content-guided video synthesis with diffusion models. *arXiv* **2023**, arXiv:2302.03011.
37. Shin, C.; Kim, H.; Lee, C.H.; Lee, S.; Yoon, S. Edit-a-video: Single video editing with object-aware consistency. *arXiv* **2023**, arXiv:2303.07945.
38. Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; Tian, Q. Controlvideo: Training-free controllable text-to-video generation. *arXiv* **2023**, arXiv:2305.13077.
39. Chai, W.; Guo, X.; Wang, G.; Lu, Y. Stablevideo: Text-driven consistency-aware diffusion video editing. In Proceedings of the IEEE/CVF International Conference on Vision, Computer, Waikoloa, HI, USA, 3–7 January 2023; pp. 23040–23050.
40. Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; Guo, B. Ccedit: Creative and controllable video editing via diffusion models. *arXiv* **2023**, arXiv:2309.16496.
41. Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. *arXiv* **2020**, arXiv:2010.02502.
42. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2022**, arXiv:2207.12598.
43. Wang, W.; Xie, K.; Liu, Z.; Chen, H.; Cao, Y.; Wang, X.; Shen, C. Zero- shot video editing using off-the-shelf image diffusion models. *arXiv* **2023**, arXiv:2303.17599.
44. Park, D.H.; Azadi, S.; Liu, X.; Darrell, T.; Rohrbach, A. Benchmark for compositional text-to-image synthesis. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), Virtual only, 6–14 December 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.