Fuse Your Latents: Video Editing with Multi-source Latent Diffusion Models

Anonymous Author(s)
Submission Id: 1490

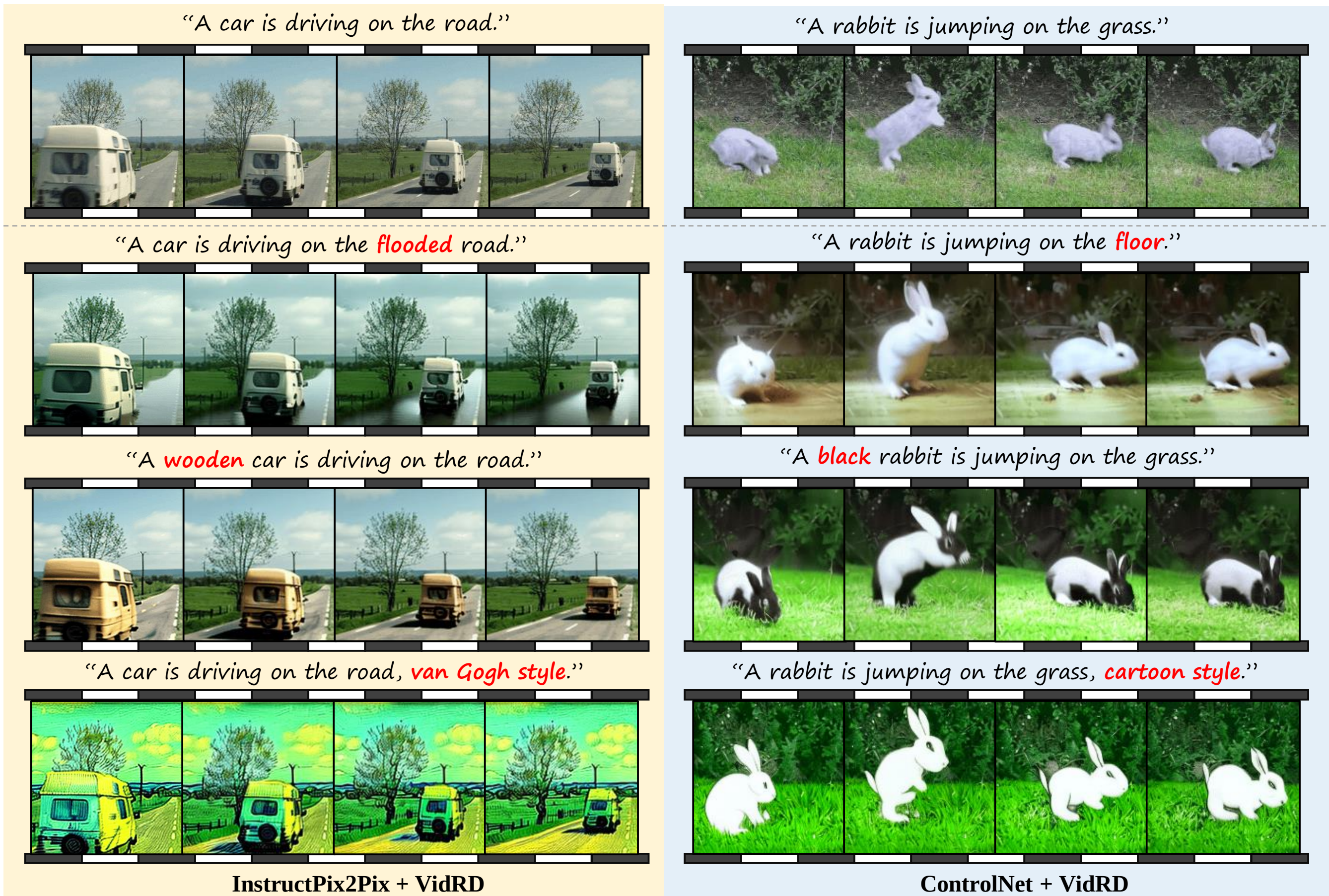

Figure 1: FLDM can serve as a versatile plugin that can be applied to off-the-shelf image diffusion models (e.g., InstructPix2Pix [3] and ControlNet [50]) and video diffusion models (e.g., VidRD [12]).

ABSTRACT

Latent Diffusion Models (LDMs) are renowned for their powerful capabilities in image and video synthesis. Yet, compared to text-to-image (T2I) editing, text-to-video (T2V) editing suffers from a lack of decent temporal consistency and structure, due to insufficient pre-training data, limited model editability, or extensive tuning costs. To address this gap, we propose FLDM (Fused Latent Diffusion Model), a training-free framework that achieves high-quality T2V editing by integrating various T2I and T2V LDMs. Specifically, FLDM utilizes a hyper-parameter with an update schedule to effectively fuse image and video latents during the denoising process. This paper is the first to reveal that T2I and T2V LDMs can complement each other in terms of structure and temporal consistency, ultimately generating high-quality videos. It is worth noting that FLDM can serve as a versatile plugin, applicable to off-the-shelf image and video LDMs, to significantly enhance the quality of video editing. Extensive quantitative and qualitative experiments on popular T2I and T2V

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

LDMs demonstrate FLDM's superior editing quality than state-of-the-art T2V editing methods. Our project page is available at https://anonymous121381.github.io/FLDM/.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Video editing, latent diffusion models, training-free framework

1 INTRODUCTION

Recently, diffusion models [16, 42] have achieved significant success in image generation [7, 10, 20, 22, 29, 37, 38, 40], image-to-image translation [39], text-to-image editing [3, 14, 18, 29] and image inpainting [21, 25, 36]. The development of text-to-image (T2I) generation and editing models has inspired text-to-video (T2V) editing. A straightforward way of T2V editing is to utilize T2I models, considering temporal consistency modeling with motion conditions [4, 6], temporal propagation [5, 11, 17, 30] and temporal modules [13, 32, 46]. Although T2I models have the advantage of video editing with high-quality structure and appearance, the direct adaptation from T2I models to T2V editing still has limitations in temporal consistency if the edited target has significant changes in shape [32] or motion [30]. Besides, the one-shot tuning strategy [1, 28, 46] suffers from low editability and training efficiency.

Therefore, another option is to develop video generation diffusion models [2, 9, 15, 44] for video editing that can achieve decent temporal consistency. Compared to the images generated by T2I models, the videos generated by T2V generation models, despite having good temporal consistency, still lack structural integrity (e.g., textures, low resolution). The main reason is that obtaining high-quality video data is more challenging compared to image data, resulting in a smaller amount of video data available for training T2V generation models [44, 45]. Additionally, to achieve high-quality video generation, T2V models require the incorporation of numerous conditional controls during training, such as depth and optical flow [9, 44], which increases the training cost compared to T2I generation models.

To this end, to achieve high-quality video editing, typically evaluated with temporal consistency, text alignment, and aesthetics, the question arises: Given T2I and T2V diffusion models, how can we maximize their potential for superior T2V editing with minimal operations? In this paper, for the first time, we investigate how to ensemble T2I and T2V diffusion models for video editing. We introduce the FLDM (Fused Latent Diffusion Model) by fusing multi-source LDMs, including T2I and T2V latent diffusion models. Drawing inspiration from Prompt2Prompt [14], certain video editing methods such as FateZero [32] facilitate local editing through attention map manipulation, which can achieve fine-grained editing results. However, this way cannot guarantee robust temporal coherence and suffers from low editability. FLDM replaces attention injection with latent fusion, which realizes an overall and global editing.

Concretely, we extract latents from both T2V and T2I LDMs. At every denoising timestep, we apply latent fusion with a hyper-parameter to control the proportion of image to video latents. The underlying intuition is that this allows us to balance temporal consistency against structural integrity. The fused latents, enriched with both temporal and semantic information from the respective models, will undergo individual denoising by each model in the subsequent timestep. This suggests that both models can leverage supplemental information—namely, temporal modeling from T2V LDMs and superior generation quality from T2I LDMs. However, the excessive guidance from T2I LDMs may introduce image artifacts during the latent fusion process. To mitigate such side effects, we adjust the fusion ratio for FLDMs during the denoising process with an update schedule. Through extensive experiments with a range of T2I and T2V LDMs, we've found that while T2V LDMs struggle to maintain the structural integerity of the source video, they are greatly enhanced by the complementary strengths of T2I LDMs. Conversely, T2V LDMs excel in ensuring robust temporal consistency in the edited videos, thereby elevating the overall quality of edits performed by T2I LDMs. Figure 1 displays several video editing results using FLDM with InstructPix2Pix [3] and ControlNet [50].

To sum up, we make the following contributions.

- We propose multi-source latent fusion, a straightforward yet effective inference strategy that can seamlessly integrate off-the-shelf T2I diffusion models with T2V diffusion models.
- To the best of our knowledge, this work is the first to reveal that T2I and T2V diffusion models can complement each other effectively. Specifically, T2I models are essential for providing structural integrity, while T2V models are crucial for ensuring temporal consistency.
- Without any tuning process, our method achieves better performance than other counterparts on both qualitative and quantitative evaluations.

2 RELATED WORK

2.1 Text-to-Image Generation

Before the advent of diffusion models, prominent text-to-image (T2I) and text-to-video (T2V) generation works predominantly adopted GAN [35, 49] or auto-regressive architectures [8, 34, 48]. Dhariwal et al. [7] were among the first to present comparisons between GANs and Diffusion Models in the context of image synthesis. Their findings suggested that diffusion models outperformed GANs in terms of diversity, stability of training objectives, and scalability. Subsequently, a series of text-to-image generation models based on the diffusion approach emerged [3, 18, 27, 29, 37, 38, 40, 50], achieving high-fidelity generation.

While T2I diffusion models have achieved remarkable performance in T2I generation, directly adapting these models for T2V editing proves insufficient due to the absence of temporal consistency modeling. Our method combines latents from both T2I and T2V models to enhance temporal consistency without compromising the editing capability of the T2I model. With no additional training or tuning required, the introduced latent fusion strategy is adaptable to any off-the-shelf T2I or T2V diffusion model, enhancing editing quality with minimal adjustments.

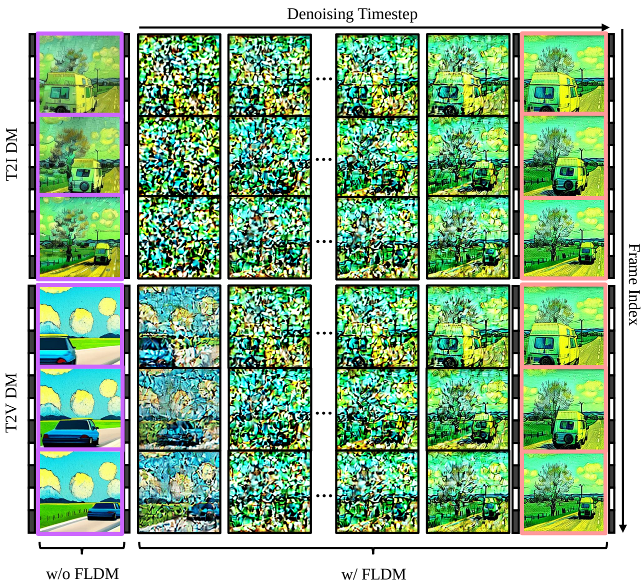

Figure 2: Denoising process of T2I and T2V LDMs (+ Van Gogh style). First column: Without FLDM, T2I LDMs have good structure preservation but lack temporal consistency, T2V LDMs lack structure preservation but achieve good temporal consistency. Last column: Both the structure and temporal consistency of the edited videos are enhanced with multi-source latent fusion (FLDM). Best viewed from project homepage.

2.2 Diffusion Models for Video Editing

A prevalent approach to video editing using diffusion models involves adapting T2I models to T2V models [41, 46]. This adaptation often incorporates temporal modules to guarantee temporal consistency. Tune-A-Video [46] integrates temporal attention layers into UNet and conducts one-shot tuning. Meanwhile, Make-A-Video [41] augments the network to encompass temporal information by extending it with spatial-temporal modules. However, fine-tuning target videos can lead to over-fitting source prompts, potentially diminishing the model's editing capabilities. To address this, several studies employ T2V diffusion models for video editing and demonstrate promising results. [28] introduces a mixed fine-tuning strategy for the Imagen Video model [15] that enhances motion editing. [9] presents a video diffusion model trained with depth information to govern video structure and content. Although these methods have yielded remarkable results in video editing, training T2V diffusion models poses a challenge due to the inherent difficulty in collecting high-quality video data for training. Furthermore, their primary focus is on enhancing T2V editing through mixed video-image fine-tuning and joint video-image training, neglecting the potential benefits of leveraging existing high-quality T2I models.

Another avenue of research is inspired by Prompt2Prompt [14] and Plug-and-Play [43], both of which facilitate local editing through attention map manipulation. [32] suggests blending self-attention maps with masks produced by cross-attention maps to support zero-shot video editing. Meanwhile, [24] introduces a decoupled-guidance attention control, adapting P2P to Video-P2P. However, these methods require extensive manual adjustment of the parameters in the attention map, making them challenging to apply in practical scenarios.

Algorithm 1 FLDM for T2V Editing

Input: Latent features from DDIM inversion by T2V and T2I model: $\mathbf{z}_T^V$, $\mathbf{z}_T^I$, target prompts P, fusion timestep τ, fusion ratio α_τ.
Output: Denoised latent $\mathbf{z}_0{}^*$

for $t = T, T-1, ..., 1$ **do**
  $\mathbf{z}_{t-1}^I \leftarrow \mathrm{DDIM}_{\mathrm{T2I}}(\mathbf{z}_t^I, P, t)$;
  $\mathbf{z}_{t-1}^V \leftarrow \mathrm{DDIM}_{\mathrm{T2V}}(\mathbf{z}_t^V, P, t)$;
  if $t \leq (T-\tau)$ **then**
    $\mathbf{z}_{t-1}^* = \alpha_t * \mathbf{z}_{t-1}^V + (1-\alpha_t) * \mathbf{z}_{t-1}^I$;
    $\alpha_{t-1} = \alpha_t + (1-\alpha_\tau)/(T-\tau)$;
    $\mathbf{z}_{t-1}^I \leftarrow \mathbf{z}_{t-1}^*$;
    $\mathbf{z}_{t-1}^V \leftarrow \mathbf{z}_{t-1}^*$;
  end if
end for
return $\mathbf{z}_0^*$

In this paper, we introduce a simple-yet-effective latent fusion strategy that harnesses the strengths of both T2I and T2V diffusion models for T2V editing without requiring any tuning. For the first time, this study demonstrates that T2I and T2V diffusion models complement each other.

3 METHOD

Our objective is to maximize the advantages of the existing text-to-image (T2I) and text-to-video (T2V) diffusion models in terms of structure integrity and temporal consistency, aiming to achieve high-quality T2V editing. Given a video $\mathcal{V}$ containing m frames, denoted as $\mathcal{V} = \{\mathbf{x}_i \mid i \in [1, m]\}$, T2V editing aims to generate a new video conditioned on an editing text prompt P. In this section, we first introduce the preliminaries of the diffusion model in Section 3.1, including latent diffusion models [37] and DDIM inversion [42]. Then the detail design of FLDM (Fused Latent Diffusion Model) will be introduced in Section 3.2 and Section 3.3.

3.1 Preliminaries

Latent Diffusion Models (LDMs). LDMs denoise noisy latents in the latent space. First, an image $\mathbf{x}$ undergoes compression via an encoder $\mathcal{E}$, producing a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$, which is reconstructed back into an image through a decoder $\mathcal{D}$. A U-Net U_θ is applied as the denoising model to predict the noise, which is optimized by minimizing the following objective

$$\min_\theta E_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t \sim \mathcal{U}(1,T)} \left\| \boldsymbol{\epsilon} - U_\theta\left(\mathbf{z}_t, t, P\right) \right\|_2^2, \tag{1}$$

where $\mathbf{z}_t$ is the noisy latent at timestep t and P is the conditional text embedding.

DDIM Inversion. Deterministic DDIM sampling inverts the noisy latent $\mathbf{z}_T$ to a clean latent $\mathbf{z}_0$, the sampling process is expressed as

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}}{\sqrt{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\boldsymbol{\epsilon}} \tag{2}$$

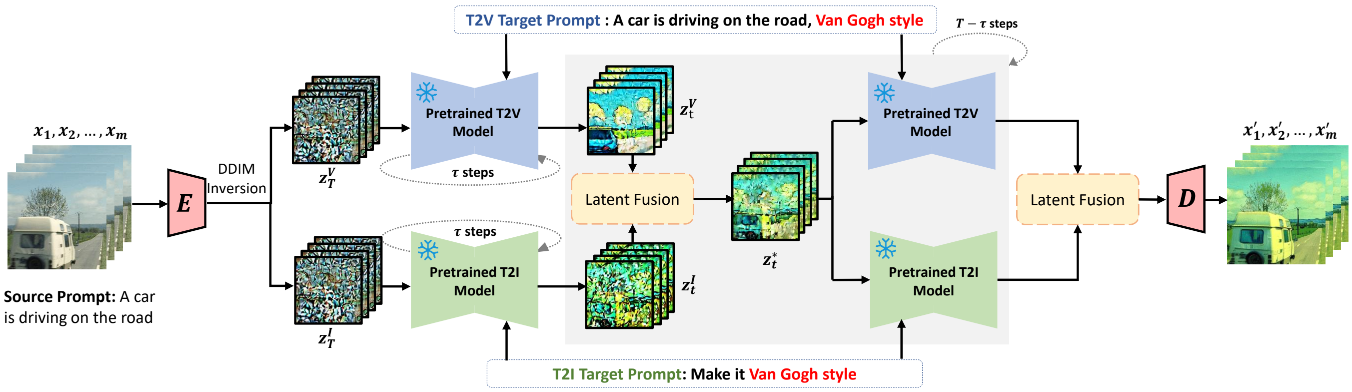

Figure 3: FLDM framework for T2V editing. During the inference stage, the input video is encoded via VAE Encoder to be a clean latent $z_0 \in \mathbb{R}^{f \times c \times h \times w}$ and then inverted to be a noisy latent $z_T \in \mathbb{R}^{f \times c \times h \times w}$ through DDIM inversion. During the first τ timesteps, the T2V LDM and T2I LDM predict noise for noisy latent respectively. In the next $T-\tau$ timesteps, a multi-source latent fusion module is applied to fuse denoised latents from T2V and T2I LDMs.

where $\hat{\boldsymbol{\epsilon}}$ is the predicted noise and α_t is the multiplication of variances at timestep t. DDIM inversion is the reverse process of DDIM sampling, it converts a clean latent $\mathbf{z}_0$ to a noised latent $\hat{\mathbf{z}}_T$ which can preserve the structure information in real images.

$$\hat{\mathbf{z}}_t = \sqrt{\bar{\alpha}_t}\frac{\hat{\mathbf{z}}_{t-1} - \sqrt{1-\bar{\alpha}_{t-1}}\hat{\boldsymbol{\epsilon}}}{\sqrt{\bar{\alpha}_{t-1}}} + \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}} \tag{3}$$

3.2 Multi-source Latent Fusion

As depicted in Figure 2, video editing conducted separately by T2I and T2V LDMs, without the integration of FLDM, reveals distinct drawbacks. T2I LDMs, although adept in frame-wise editing, tend to cause flickering issues, e.g., the tree in the background has obvious shape changes across frames. Similarly, T2V LDMs, despite their ability to maintain temporal consistency, often compromise on structural quality, e.g., the car shape is different from the original car (see the first row in Figure 1). FLDM addresses these issues by strategically fusing the latents of both T2I and T2V LDMs during the denoising process. As the denoising timesteps progress, there is an improvement in both the structural integrity and temporal consistency of the video edits.

Figure 3 shows the FLDM pipeline of video editing with multi-source latent fusion. Given the source video $\mathcal{V}_{\text{src}}$ and a target prompt P, the VAE Encoder compresses the input video into a clean latent $\mathbf{z}_0 \in \mathbb{R}^{f \times c \times h \times w}$, where f represents the number of frames in the source video. Then we apply DDIM inversion with T2V and T2I LDMs, which makes use of the source prompt, to invert the clean latent into noisy latents $\mathbf{z}_T^V \in \mathbb{R}^{f \times c \times h \times w}$ and $\mathbf{z}_T^I \in \mathbb{R}^{f \times c \times h \times w}$. Afterward, T2V and T2I LDMs are applied to predict noise. Note that the T2I LDMs predict noise frame-by-frame and the denoised T2I latents are concatenated at the temporal dimension. Then we apply latent fusion by combining the denoised latents from T2V and T2I LDMs. As a result, a mixed noisy latent $\mathbf{z}_t^*$ is generated for the next denoising timestep

$$\mathbf{z}_t^* = \alpha_t \mathbf{z}_t^V + (1-\alpha_t)\mathbf{z}_t^I \tag{4}$$

where the hyper-parameter α controls the ratio of T2V and T2I latents. A larger α ensures better temporal consistency but does harm to structure and text alignment, while if α is too small, it would degrade to frame-wise editing with T2I LDMs.

Notice that, our method is a versatile plugin that can work for latent diffusion models within the same latent space. We need to ensure that two LDMs share the same VAE Encoder and Decoder. Most LDMs are trained on the basis of Stable Diffusion [37], where the weights of VAE are frozen and remain unchanged. As we utilize off-the-shelf T2I LDMs and T2V LDMs for joint denoising, the editing results depend on the fused models. Specifically, the temporal modeling capability of T2V LDMs affects the temporal consistency of edited video frames, while the spatial structure of the videos is related to T2I LDMs.

3.3 Fusion Ratio Update Schedule

In practice, we notice that fusing multiple latents at the early diffusion steps results in noisy output, since latent fusion may break the structure of videos. As a result, we let T2V and T2I LDMs infer separately in early τ steps and perform latents fusion at afterward steps. Although α balances the ratio of temporal consistency and structure modeling, a fixed α may bring too much guidance from T2I LDMs at the end of the denoising process, thus causing temporal inconsistency in generated videos. Since T2V LDMs are able to generate temporal coherent videos, we increase the ratio of T2V latents at the end of the denoising process. We propose an update schedule for the fusion ratio, wherein the ratio is incrementally increased after each latent fusion step as following:

$$\alpha_{t-1} = \alpha_t + (1-\alpha_\tau)/(T-\tau) \tag{5}$$

where T is the total number of denoising steps, α_t is the fusion ratio at timestep t. In practice, we start to increase α_t at timestep τ. The initial fusion ratio α_τ ranges between [0, 1]. Formally, the algorithm is shown in Algorithm 1.

FLDM aims to augment the capability of T2V models through T2I models, ensuring that T2V models retain a prominent position. So T2V latents should be allocated a larger propotion. The T2V

Figure 4: Qualitative comparison with SOTA approaches. FLDM has the best textual alignment, temporal consistency, and fidelity.

signals and T2I signals in fused latents can be calculated as:

$$\mathcal{S}_V = \int_{\tau}^{T} \alpha_t dt \tag{6}$$

$$\mathcal{S}_I = \int_{\tau}^{T} (1 - \alpha_t) dt \tag{7}$$

As α_t changes in a linear manner, we can calculate the difference between the T2V and T2I contributions:

$$\delta = \mathcal{S}_V - \mathcal{S}_I = 2\alpha_\tau (T - \tau) > 0 \tag{8}$$

which means T2V signals take a dominant role. On the other hand, a linear update schedule ensures a smoother denoising process than a polynomial way.

4 EXPERIMENTS

4.1 Experiment Settings

Implementation Details. We attempt to apply FLDM to various T2I and T2V diffusion models to validate the effectiveness. For the diffusion model gallery, we consider popular T2I diffusion models including ControlNet [50] and InstructPix2Pix [3] loaded with pre-trained weights from Stable Diffusion v1.5 [37]. Although there are some high-quality video diffusion models [9, 15, 45], unfortunately, they can not be accessed. Therefore, we take the publicly available VidRD [12] and ZeroScope [26] with released pre-trained weights. We then sample 8 frames uniformly from input videos with a resolution of 256p for all models. For T2V models fused with ControlNet, we use ControlNet with canny edge condition and set the total timestep $T = 50$ for DDIM schedule. For T2V models fused with InstructPix2Pix, we set the total timestep $T = 100$ and the classifier-free guidance scale of text to 12.5 and of image to 1.5. We select 16 videos from DAVIS dataset [31] for evaluation, which are part of the TGVE benchmark [47]. For each video, there are three types of text prompts for video editing including *style*, *object*, and *background*.

Evaluation Metrics. Following previous text-to-video editing works [9, 32, 46], we conduct both quantitative and qualitative evaluations for FLDM. For quantitative evaluation, we utilize automatic metrics including frame consistency (**'Tem-Con'**), text alignment (**'Text-Align'**), user preference (**'User-Pre'**) with CLIPScore [33] and PickScore [19]. For user study, three metrics (denoted as **'Edit'**, **'Image'**, and **'Temp'**) are conducted to evaluate editing quality, frame fidelity, and temporal consistency of edited video. Following [23], we recruit 20 evaluators to pair-wisely compare our method with baselines, and present the preference percentage with "p_1/p_2" where p_1 denotes the preference percentage of baseline, p_2 denotes the preference percentage of our method.

4.2 Main Results

Baselines. We test two T2I models ControlNet (CN) and InstructPix2Pix (IP2P), and two T2V models VidRD and ZeroScope (ZS) under the individual testing setting (T2I or T2V) and FLDMs testing setting (T2I + T2V). In addition, we compare our method against two

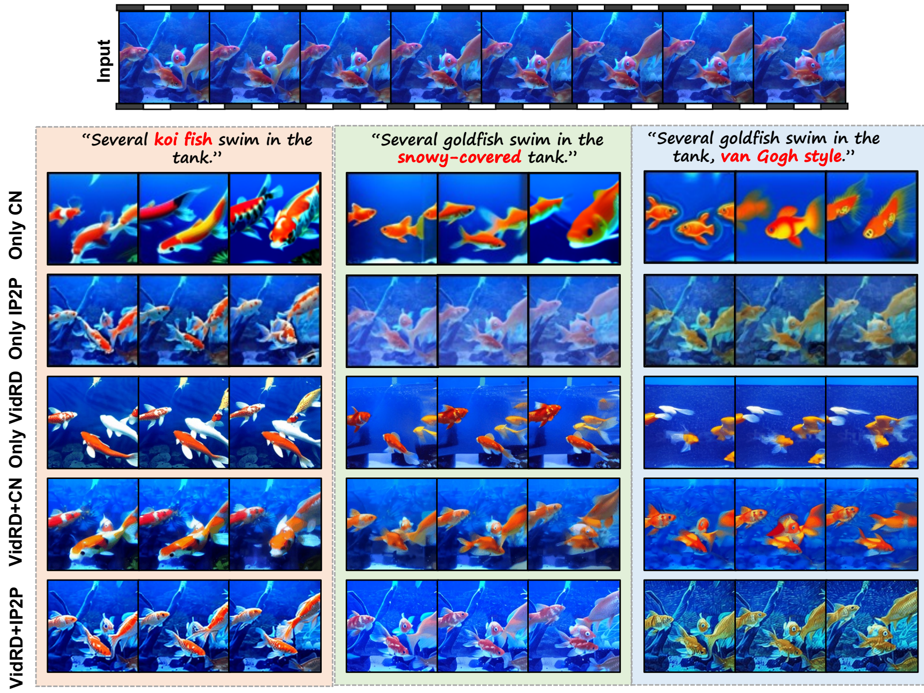

Figure 5: Object, background, and style editing results of FLDM. In comparison with T2I models only, FLDM with VidRD generates more consistent frames. Compared to videos edited by VidRD only, FLDM is superior in structure and fidelity.

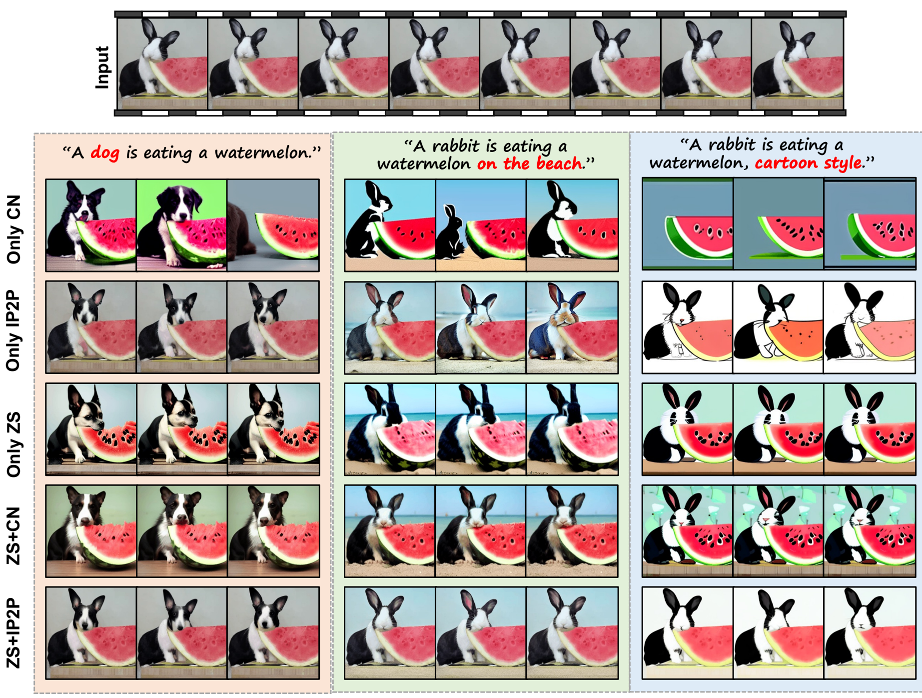

Figure 6: Results of object, background, and style editing using FLDM with ZeroScope. FLDM is effective for improving editing fidelity and is flexible for various T2I and T2V diffusion models.

Table 1: Automatic evaluation of FLDMs and other baseline methods. FLDMs achieve the best textual alignment, user preference, and comparable temporal consistency.

Method Inversion & Editing	CLIP Mertics Tem-Con	 Text-Align	PickScore User-Pre
Framewise IP2P [3]	86.76	25.11	20.33
Framewise CN & DDIM [50]	80.02	25.41	20.12
Tune-A-Video & DDIM [46]	90.45	27.13	20.42
FateZero [32]	**92.92**	23.81	20.22
VidRD & DDIM [12]	92.03	27.18	20.58
ZS & DDIM [26]	90.24	27.49	20.58
FLDMs			
VidRD + IP2P	89.28	26.71	20.66
VidRD + CN	90.14	27.43	20.64
ZS + IP2P	90.12	27.50	**20.68**
ZS + CN	88.24	**27.70**	20.60

Table 2: User preference of FLDMs and other baseline methods. FLDMs achieve the highest human preference over all evaluation metrics and outperform all baselines by a clear margin.

Method Inversion & Editing	 Edit	User Study Image	 Temp
Framewise IP2P [3]	31.60 / **68.40**	47.90 / **52.10**	5.25 / **94.75**
Framewise CN & DDIM [50]	21.10 / **78.90**	10.50/ **89.50**	10.50 / **89.50**
Tune-A-Video & DDIM [46]	33.32 / **66.68**	20.18 / **79.82**	18.43 / **81.57**
FateZero [32]	35.10 / **64.90**	39.63 / **60.37**	35.10 / **64.90**
VidRD & DDIM [12]	32.65 / **67.35**	5.25 / **94.75**	35.80 / **64.20**
ZS & DDIM	34.75 / **65.25**	39.50 / **60.50**	40.25 / **59.75**

state-of-the-art video editing methods including Tune-A-Video [46] and FateZero [32].

Applications. We show three example applications including 1) *Object editing*: Our method has a significant use case in altering objects by manipulating text prompts. This enables effortless replacement, addition, or removal of objects. For example, we can replace "goldfish" with "koi fish" or "sharks" (see Figure 5). When using our method with InstructPix2Pix, the instruction can be "replace the goldfish with koi fish" or just "turn them sharks". 2) *Background change*: Another application of our method is to change the background of video. When altering the location of the object, our method maintains the coherence of the object's motion. For example, we can modify the background of the rabbit in Figure 1 to be "on the floor", and add shadow of the rabbit that doesn't exist in the origin video. Figure 5 shows the results of editing "tank" to be "snowy-covered tank". 3) *Style transfer*: Through taking the knowledge from T2I models in open domains, FLDM can facilitate transforming videos into diverse styles that are challenging to acquire solely from video data. For example, we transform real-world videos into cartoon style (Figure 1), or Van Gogh style (Figure 5), by adding style descriptors to the prompt.

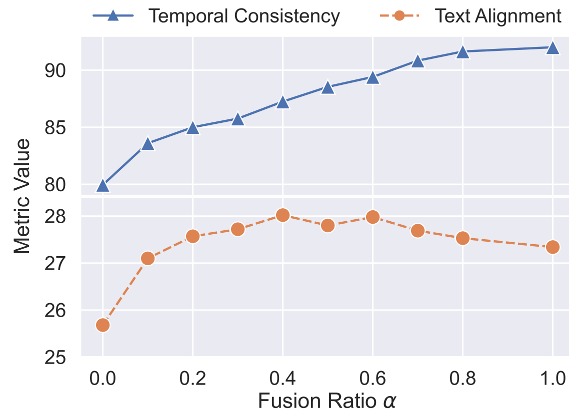

Figure 7: Temporal consistency and text alignment result with various fusion ratios α. Temporal consistency can be improved with larger α, and T2V and T2I models complement each other in textual alignment.

Qualitative Results. We showcase visual comparisons of our method against other baselines from Figure 4 to Figure 6. In Figure 5 and Figure 6, we present the effects of fusing T2I models with VidRD and ZeroScope separately. In the case of goldfish, InstructPix2Pix has superior capabilities of frame-wise editing, while lacking consistency between frames (e.g., the appearance of koi fishes exhibits significant variations). In contrast, videos generated by VidRD individually have decent temporal consistency but lack original structure (e.g., background details). When applying FLDM to VidRD and InstructPix2Pix, both temporal consistency and structural integrity can be improved. In Figure 6, we observed that ZeroScope (ZS) achieves satisfactory video editing results with DDIM inversion. However, FLDM can further enhance the visual aesthetics and fidelity of the generated videos (e.g., the cartoon rabbit generated by FLDM is more similar to the original rabbit). In Figure 4, we compare FLDMs with two state-of-art video editing methods. We find Tune-A-Video struggles to keep the fidelity of the original video (e.g., the car is different from the one in the original video) since it lacks regional control. Besides, videos generated by FateZero fail to align with target prompts well (e.g., Vermeer painting style) without carefully tuning on word strength hyper-parameters. It is worth noting that FLDM is more efficient than SOTAs since it does not require any model tuning [46] or heavy handcrafted hyper-parameter tuning [32].

Quantitative Results. We use automatic metrics to quantify our method against baselines in Table 1. Results indicate that compared to editing with T2V models only, FLDM achieves higher user preference and better textual alignment at the cost of a slight loss of temporal consistency. FateZero demonstrates good frame consistency but struggles to achieve accurate text-guided editing. Tune-A-video excels in both temporal consistency and textual alignment. In comparison, our method outperforms these two methods in textual alignment and achieves comparable temporal consistency. The user study results are shown in Table 2, which indicates our method is preferred by users across all metrics.

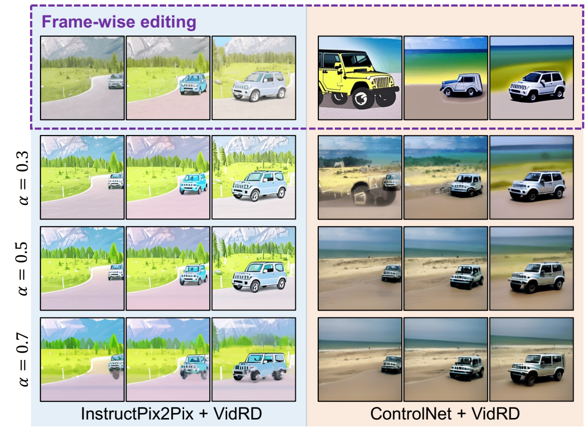

Figure 8: Ablation of Initial Fusion Ratio α. The editing prompts are "*a jeep car is moving on the road, cartoon style.*" and "*a jeep car is moving on the beach.*".

4.3 Ablation Study

For all the ablation experiments, we take VidRD as the T2V model, which is fused with T2I models (ControlNet and InstructPix2Pix) through FLDM if not other mentioned.

Ablation of Initial Fusion Ratio α. We conduct extensive experiments with various fusion ratios α as shown in Figure 7 and have the following observations: 1) For temporal consistency, the larger α makes T2V latents have larger weight in fused latents which advances temporal consistency for generated videos. 2) For text alignment, there is one optimal value between $[0, 1]$ which indicates that T2V and T2I models are complementary for text alignment, probably because the T2V model is trained with large-scale video datasets that have rich action concepts in complementary with T2I text prompts [45]. Figure 8 shows some samples generated with two T2I architectures fused with the T2V model. As we can see, the frame-wise editing (only T2I) performs worst since it ignores the temporal relation among frames. While with a T2V model and a suitable fusion ratio α, the temporal consistency can be improved clearly. Since training a T2V diffusion model is expensive in collecting datasets and computational cost, we can make the most use of existing T2I models to enhance T2V editing with FLDM. Considering the trade-off between temporal consistency and text alignment, we fuse a smaller proportion of T2I latents compared to T2V latents in practice.

Ablation of τ. Table 3 shows the quantitative results of different τ. Figure 9 illustrates the effectiveness of hyper-parameter τ which decides from which denoising timestep we start to perform FLDM. When FLDM starts at an early timestep (e.g., $\tau <= 30$ for InstructPix2Pix), it may destroy the edited video structure since T2V latents are incorporated into the denoising process too early. If FLDM starts too late (e.g., $\tau >= 45$ for ControlNet), the T2I latents would take too much proportion throughout the denoising process, as a result, destroy temporal consistency for the edited video. As shown in Figure 9, the car in the bottom row exhibits artifacts.

Ablation of α Update Schedule As we can see from Figure 10, the α update schedule can improve temporal consistency, e.g., the appearance of the car (Left: InstructPix2Pix + VidRD) and the beach (Right: ControlNet + VidRD) can be better preserved among video frames with the update schedule. This illustrates that, with decayed T2I latent proportion at the last $T - \tau$ timesteps, the update schedule further reduces T2I latents' negative effect on temporal consistency.

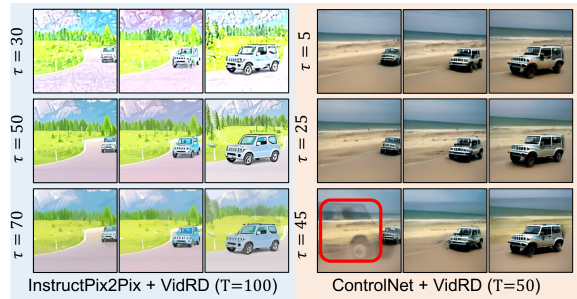

Figure 9: Effects of τ, the timestep that FLDM starts to fuse T2V latents with T2I latents during the denoising process.

Table 3: Ablation study of different τ.

Method	timestep(τ)	CLIP Mertics		PickScore
		Tem-Con	Text-Align	User-Pre
VidRD + IP2P	30	88.73	24.89	20.42
	50	**89.09**	**24.95**	**20.49**
	70	88.11	24.79	20.43
VidRD + CN	5	86.59	**27.55**	20.60
	25	**90.14**	27.43	**20.64**
	45	82.60	25.41	20.14

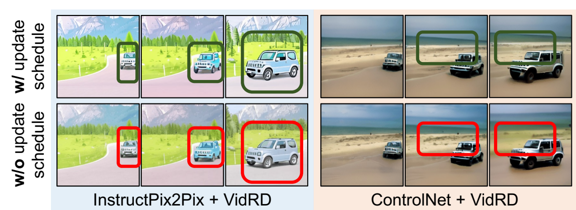

Figure 10: Effectiveness of update schedule for fusion ratio α, examples show it can improve temporal consistency while reducing image artifacts.

5 CONCLUSION

In this paper, we propose FLDM (Fused Latent Diffusion Model), a simple-yet-effective strategy that achieves high video editing quality without any tuning cost. For each denoising timestep, we apply a hyper-parameter to adjust the latent ratio of T2V and T2I diffusion models, with an update schedule to alleviate image artifacts. For the first time, we reveal that T2V and T2I diffusion models are complementary to each other in temporal consistency and structure. Our method can serve as a versatile plugin for various off-the-shelf T2I and T2V models, which we believe will be valuable for real-world practice.

REFERENCES

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

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22563–22575.

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

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

[5] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23040–23050.

[6] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. 2023. Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2305.13840* (2023).

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021), 19822–19835.

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011* (2023).

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

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373* (2023).

[12] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. 2023. Reuse and Diffuse: Iterative Denoising for Text-to-Video Generation. *arXiv preprint arXiv:2309.03549* (2023).

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

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

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

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

[17] Jiahui Huang, Leonid Sigal, Kwang Moo Yi, Oliver Wang, and Joon-Young Lee. 2023. Inve: Interactive neural video editing. *arXiv preprint arXiv:2307.07663* (2023).

[18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[19] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation.

[20] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

[21] Wenbo Li, Xin Yu, Kun Zhou, Yibing Song, Zhe Lin, and Jiaya Jia. 2022. Sdm: Spatial diffusion model for large hole image inpainting. *arXiv preprint arXiv:2212.02963* (2022).

[22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

[23] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. 2023. DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing. arXiv:2310.10624 [cs.CV]

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

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.

[26] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. arXiv:2303.08320 [cs.CV]

[27] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

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

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[30] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. 2023. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926* (2023).

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

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

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

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[36] MMOBB Richard and MYS Chang. 2001. Fast digital image inpainting. In *Appeared in the Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP 2001), Marbella, Spain*. 106–107.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*. 22500–22510.

[39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

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

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

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[43] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. 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*. 1921–1930.

[44] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018* (2023).

[45] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. 2023. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103* (2023).

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

[47] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, Rui He, Feng Hu, Junhua Hu, Hai Huang, Hanyu Zhu, Xu Cheng, Jie Tang, Mike Zheng Shou, Kurt Keutzer, and Forrest Iandola. 2023. CVPR 2023 Text Guided Video Editing Competition. arXiv:2310.16003 [cs.CV]

[48] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[49] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.

[50] Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).