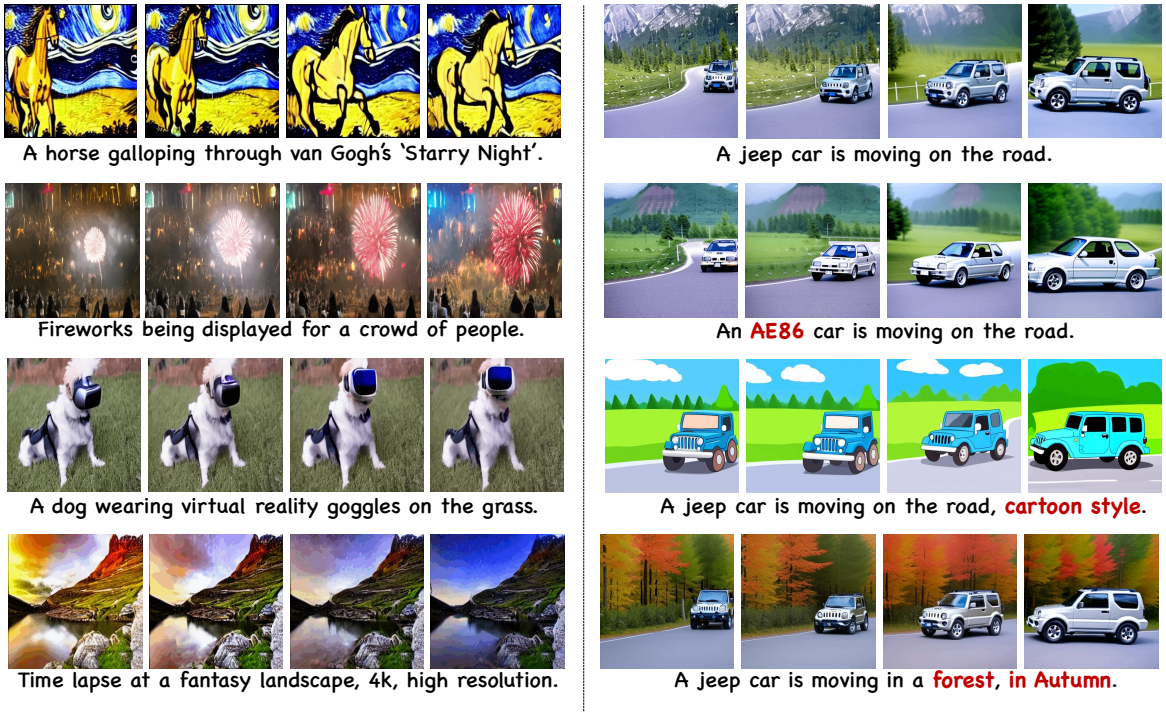
SimDA: Simple Diffusion Adapter for Efficient Video Generation

Zhen Xing[1,2] Qi Dai[3] Han Hu[3] Zuxuan Wu[1,2†] Yu-Gang Jiang[1,2]
[1] Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3] Microsoft Research Asia
https://chenhsing.github.io/SimDA



(a) Text-to-Video Generation (b) Text-guided Video Editing

Figure 1. ***Examples of our SimDA results:*** (a) The results of open-wild Text-to-Video Generation. (b) Text-guided video Editing results using one text-video pair tuning.

Abstract

The recent wave of AI-generated content has witnessed the great development and success of Text-to-Image (T2I) technologies. By contrast, Text-to-Video (T2V) still falls short of expectations though attracting increasing interest. Existing works either train from scratch or adapt large T2I model to videos, both of which are computation and resource expensive. In this work, we propose a Simple Diffusion Adapter (SimDA) that fine-tunes only 24M out of 1.1B parameters of a strong T2I model, adapting it to video generation in a parameter-efficient way. In particular, we turn the T2I model for T2V by designing light-weight spatial and temporal adapters for transfer learning. Besides, we change the original spatial attention to the proposed Latent-Shift Attention (LSA) for temporal consistency. With a similar model architecture, we further train a video super-resolution model to generate high-definition (1024×1024) videos. In addition to T2V generation in the wild, SimDA could also be utilized in one-shot video editing with only 2 minutes tuning. Doing so, our method could minimize the training effort with extremely few tunable parameters for model adaptation.

† Corresponding author.

1. Introduction

Image generation stands on top of the recent AIGC wave. It not only has a significant impact on the academic community but also achieves tremendous success in various applications, such as computer graphics, art and culture, medical imaging, *etc*. The approaches in this area mainly include methods based on generative adversarial networks (GANs) [29, 49–51, 84], auto-regressive transformers [20, 77, 125], and the latest diffusion models [17, 23, 38, 40, 58, 67, 68, 78, 79, 82, 91, 120]. Among them, diffusion models are the most popular owing to their strong controllability, simple stability, and amazing realism. However, video generation research lags behind due to challenges like the scarcity of publicly available datasets, difficulty in modeling temporal information, and high training costs, hindering progress in this area.

There have been several research endeavors dedicated to exploring video synthesis [3, 9, 12, 15, 24, 27, 30, 31, 42, 48, 52, 55, 60, 62, 64, 70, 83, 89, 94, 100, 102, 107, 108, 121, 126]. In addition, some studies have employed popular diffusion models for video generation [32, 41, 43, 61, 101, 106, 116, 122, 135, 137]. However, most of them involve training models from scratch, which can be time-consuming due to the complex video data. Early attempts were also constrained by GPU memory or hardware limitations.

More recently, a small number of T2V (Text-to-Video) approaches have emerged, aiming to fine-tune well-established T2I (Text-to-Image) models [1, 8, 28, 39, 61, 104, 136]. They have incorporated temporal modeling modules (*e.g.* Imagen video [39], Video LDM [8]) into T2I models, which effectively accelerate the model convergence. However, it should be noted that training such models is still a challenging task due to the massive number of parameters (4B or even 16B) involved in the network architecture.

In the NLP field, state-of-the-art results of various tasks are generally achieved by adaptation from large pretrained models (*i.e.*, BERT [16], LLMs [14, 74, 76, 132]). However, with the advent of increasingly larger and more powerful foundation models (*e.g.*, GPT-4 with 100T parameters), conducting full fine-tuning of the entire models has become prohibitively expensive and infeasible in terms of training cost and GPU storage. To address the issue, numerous methods based on efficient fine-tuning have emerged rapidly in NLP [44, 45, 53, 54] and computer vision [13, 69, 124, 128].

In this work, we propose a parameter-efficient video diffusion model, namely Simple Diffusion Adapter (SimDA), that fine-tunes the large T2I (*i.e.* Stable Diffusion [79]) model for improved video generation. We only add 2% parameters compared to the T2I model. During training, we freeze the original T2I model, and only tune the newly added modules. We further propose a Latent-Shift Attention (LSA) to replace the original spatial attention, which significantly improves the temporal modeling capability and retains consistency without adding new parameters. To this end, our model demands less than 8GB GPU memory for training with a resolution of $16 \times 256 \times 256$, while the inference time speeds up by $39\times$ compared to the auto-regressive method CogVideo [42]. Besides, we turn an image super resolution framework into the video counterpart with similar architecture, which allows generating high-definition videos of 1024×1024. Our model can also be extended to the recently popular diffusion-based video editing [110], achieving significant $3\times$ faster training while retaining comparable results, as evidenced by the editing examples presented in Fig 1 (b). In conclusion, the contributions of this work can be summarized as follows:

- We explore the simple adaptation from image diffusion to video diffusion, exhibiting that tuning extremely few parameters can achieve surprisingly good results.
- With the helpful light-weight adapters and the proposed latent-shift attention, our method can effectively model the temporal relations with negligible cost.
- Our diffusion adapter could be extended to text-guided video super-resolution and video editing, significantly facilitating the model training.
- SimDA significantly alleviates the training cost and speeds up the inference time, while remaining competitive results compared to other methods.

2. Related Work

Text-to-Video Generation Similar to the advancements in Text-to-Image (T2I) generation [18, 71, 79, 130], early approaches for Text-to-Video (T2V) generation [55, 64, 70] were based on Generative Adversarial Networks (GANs) and primarily applied to domain-specific videos such as simple human actions [92] or clouds moving [117].

Recently, T2V methods are most based on fine tuning the T2I [78, 79] diffusion models. For instance, Make-A-Video [88] proposes fine-tuning a pretrained DALLE2 [78] model solely on video data to learn motion patterns, enabling T2V generation without explicitly training on text-video pairs. Video Diffusion Models [41] and Imagen Video [39] perform joint text-image and text-video training, treating images as independent frames and disabling temporal layers in the U-Net [80] architecture. Besides, Video LDM [8], Latent-Shift [1], VideoFactory [104], MagicVideo [136] and our methods utilize the popular open-sourced T2I Stable Diffusion [79] model. While the progress in video generation is impressive, the parameters of video generation can be highly large. As shown in Table 1, Make-A-Video [88] requires six models and 9.7B parameters and Imagen Video [39] utilizes eight models with 16.3B parameters, which limits the training efficiency of T2V models.

Text guided Video Editing In the realm of content gen-

eration, an alternative avenue is the manipulation of existing images [10, 37, 63, 98] and videos [6, 21, 59, 73, 87, 110, 115, 123, 134] using textual input as a means of control, rather than relying solely on unbridled text-based generation. SDEdit [63] introduces noise to images and then reconstructs them for the purpose of editing. Prompt-to-prompt [37] and Plug-and-Play [98] modify the cross-attention map by altering the textual description, thus influencing the editing process. When it comes to video editing, Tune-A-Video [110] fine-tunes the T2I model on a single video, enabling the generation of new videos with similar motion patterns. Video-P2P [59] and FateZero [73] extend the concept of Prompt-to-prompt editing to videos. Text2Live [6] divides videos into layers and enables separate editing of each layer based on text. MotionEditor [97] edits the motion of human while keeping the background.

Parameter-Efficient Transfer Learning In NLP, parameter efficient fine-tuning [33, 44, 45, 53, 54, 93, 127] were initially proposed to address the heavy computation of full fine-tuning LLMs for various downstream tasks. These techniques aim to reduce the number of trainable parameters, thereby lowering computation costs, while still achieving or surpassing the performance of full fine-tuning. Recently, parameter-efficient transfer learning has also been explored in the field of computer vision [4, 13, 22, 26, 46, 47, 69, 95, 112, 113, 124, 128]. These methods mainly focus on adapting models within simple classification or detection tasks. In contrast, our approach focuses on adapting a T2I model for T2V generation task.

Temporal Shift Module TSM [56] pioneered the introduction of the temporal shift module for action recognition, employing a partial channel shift along the temporal dimension. This approach seamlessly integrates temporal cues from both preceding and succeeding frames into the current frame without incurring additional computational overhead. Subsequently, TokShift [129] implemented channel shifting along the temporal dimension for transformer architectures [19, 133]. TPS [111] further shifted patches instead of channels to model the temporal correlations. However, such direct patch shifting would lead to inconsistency in generation tasks. Additionally, Latent-shift [1] and TSB [66] adapted shift module as TSM [56] within convolution blocks for video generation tasks. In this work, our latent-shift attention (LSA) employs the patch-level shifting manner. In contrast to TPS, we further propose to involve all tokens in the current frame as the keys and values, which guarantees temporal consistency during generation and significantly improves the video quality.

3. Method

In this section, we first introduce the preliminaries of Latent Diffusion Model [79] in Sec. 3.1. The pipeline of SimDA is described in Sec. 3.2. Then we detail the proposed spatial and temporal adapters as well as latent-shift attention in Sec. 3.3. Finally, we introduce the super resolution and text-guided video editing model in Sec. 3.4.

3.1. Preliminaries of Stable Diffusion

In this subsection, we introduce the preliminaries of Stable Diffusion [79] model. It is a latent diffusion model that operates in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where $\mathcal{E}$ is the encoder and $\mathcal{D}$ is the decoder. In this model, for an image I with its corresponding latent feature $\boldsymbol{x}_0 = \mathcal{E}(I)$, the diffusion forward process involves iteratively adding noise to the latent space,

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (1)$$

where $t \in \{1, ..., T\}$ is the time step, $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is the conditional density of $\boldsymbol{x}_t$ given $\boldsymbol{x}_{t-1}$, $\mathbf{I}$ is identity matrix, and α_t is hyperparameter. Alternatively, we can directly sample $\boldsymbol{x}_t$ at any time step from $\boldsymbol{x}_0$ with,

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad (2)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

In the diffusion backward process, a U-Net denoted as $\boldsymbol{\epsilon}_\theta$ is trained to predict the noise in the latent space, aiming to iteratively recover $\boldsymbol{x}_0$ from $\boldsymbol{x}_T$. In this process, as the diffusion progresses and approaches a large value of T, $\boldsymbol{x}_0$ is completely disrupted and the latent representation $\boldsymbol{x}_T$ approximates a standard Gaussian distribution. Consequently, the U-Net $\boldsymbol{\epsilon}_\theta$ is trained to infer meaningful and valid $\boldsymbol{x}_0$ from random Gaussian noises. The training object can be simplified as,

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t)||_2^2], \quad (3)$$

where $\boldsymbol{c}$ is the embedding of condition text.

During the inference stage, it samples a valid latent representation $\boldsymbol{x}_0$ from the standard Gaussian noise $\boldsymbol{x}_T = \boldsymbol{z}_T, \boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using DDIM [90] sampling. Then, the model can decode $\boldsymbol{x}_0$ using the decoder $\mathcal{D}$ to generate the final image $I = \mathcal{D}(\boldsymbol{x}_0)$. This process could generate diverse and high-quality images based on the sampled latent representations. In contrast, our method focuses on more challenging high-quality video generation.

3.2. Pipeline

SimDA, as shown in Fig. 2, is built upon the previously introduced Stable Diffusion [79]. For a video clip with t frames, denoted as $\{I_i\}_{i=1}^t$, we first pass it through a pre-trained encoder $\mathcal{E}$ to obtain the corresponding latent feature $\{\boldsymbol{x}_i\}_{i=1}^t$. We then input the latent features to the forward diffusion process, where noise is incrementally added to the latents. In the backward diffusion process, we utilize an inflated U-Net architecture to predict the noise for the

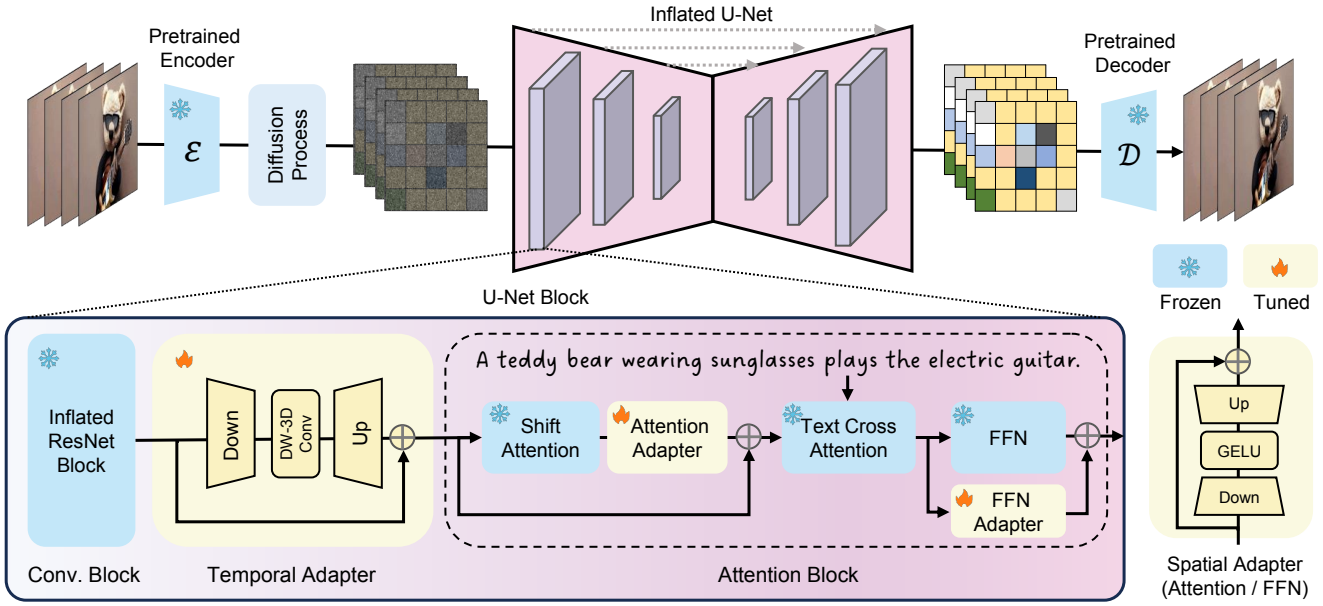


Figure 2. Pipeline of our Parameter-Efficient Text-to-Video Framework. We utilize the pre-trained auto-encoder as in Stable Diffusion [79] to obtain latent representation. During training, we only update the parameters of the newly added adapter module, highlighted in yellow. The parameters of other modules are frozen, highlighted in blue.

Table 1. Model size and inference speed comparisons. The speed is measured in seconds on one A100 (80GB) GPU. The majority of results are sourced from [1].

Method	Parameters (Billion)								Speed (s)
	T2V Core	Auto Encoder	Text Encoder	Prior Model	Super Resolution	Frame Interpolation	Overall	Tuned	
CogVideo [42]	7.7	0.10	–	–	–	7.7	15.5	15.5	434.53
Make-A-Video [88]	3.1	–	0.12	1.3	$1.4+0.7$	3.1	9.72	9.72	–
Imagen Video [39]	5.6	–	4.6	–	$1.2+1.4+0.34$	$1.7+0.78+0.63$	16.25	16.25	–
Video LDM [8]	1.51	0.08	0.12	–	0.98	1.51	4.20	2.65	–
Latent-VDM [1]	0.92	0.08	0.58	–	–	–	1.58	0.92	28.62
Latent-Shift [1]	0.88	0.08	0.58	–	–	–	1.53	0.88	23.40
LVDM [35]	0.96	0.08	0.12	–	–	–	1.16	1.04	21.23
SimDA (Ours)	0.88	0.08	0.12	–	–	–	**1.08**	**0.025**	**11.20**

noisy video latents. Specifically, for the Convolution block, we inflate the 2D ResNet [34] block to a 3D block to accommodate video inputs. Additionally, we incorporate a lightweight Temporal Adapter module for temporal modeling. In the Attention block, we employ a latent-shift attention mechanism for spatial self-attention and introduce two spatial adapter modules to facilitate the transfer of video information. Further details are presented in Sec. 3.3. During inference, we employ DDIM [90] sampling to progressively denoise the latent representation sampled from a standard Gaussian distribution. Finally, we utilize a pre-trained decoder $\mathcal{D}$ to reconstruct the video from the denoised latent.

3.3. Modeling

In this section, we describe the proposed Spatial Adapter, Temporal Adapter, and Latent-Shift Attention in detail, which are the key components of our model.

Spatial Adapter The large-scale text-image pre-trained T2I model exhibits significant transferability, as evidenced by its remarkable accomplishments in tasks such as personalized T2I generation [65, 81] and image editing [37, 131]. Consequently, we believe that employing a lightweight fine-tuning approach can effectively harness spatial information in the realm of video generation. Inspired by efficient fine-tuning techniques in NLP [45, 54] and vision tasks [13, 124], we adopt adapters due to their simplicity.

In our T2V framework, we design two types of spatial adapters (*i.e.*, Attention Adapter and FFN Adapter) for transferring video spatial information. As shown in the bottom right of Fig. 2, both adapters employ a bottleneck architecture consisting of two fully connected (FC) layers with an intermediate activation layer. The first FC layer maps the input to a lower-dimensional space, while the second FC layer maps it back to the original dimensional. Formally, for an input feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, the spatial adapter could be written as:

$$\texttt{S-Adapter}(\mathbf{X}) = \mathbf{X} + \mathbf{W}_{\text{up}}(\texttt{GELU}(\mathbf{W}_{\text{down}}(\mathbf{X}))), \quad (4)$$

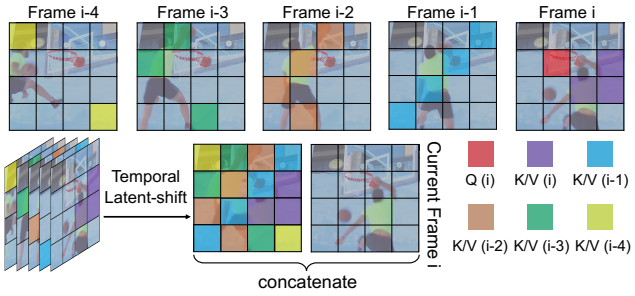


Figure 3. The overview of Temporal Latent-shift Attention module. It is noted that the Latent-shift attention is performed on latent space, but the visualization overview is shown on the image level for clear presentation.

where $\mathbf{W}_{\texttt{up}}$ and $\mathbf{W}_{\texttt{down}}$ are the learnable matrix with dimension $d \times l$ and $l \times d$, $l < d$. To preserve the structure of the original network and the pretrained weights, we initialize the second FC layer $\mathbf{W}_{\texttt{down}}$ with zeros. To adapt to the spatial features of videos, we incorporate the adapter after the latent-shift attention layer. Additionally, we observe that adding an adapter to the feed-forward network (FFN) also helps the network transfer spatial information to videos. We will provide examples in Sec. 4.4. During training, all layers of the attention block are fixed, and only the adapters are updated.

Temporal Adapter While the spatial adapter effectively transfers spatial information to video, modeling temporal information is crucial for T2V generation tasks. Previous approaches incorporate temporal convolution [11, 85, 86] or temporal attention [8, 88, 104] modules to capture temporal relationships. Although these modules are effective in modeling temporal dynamics, they often come with a huge number of parameters and high-dimensional input feature, resulting in significant computational and training costs.

To address this issue, we utilize the temporal adapter module for temporal modeling as [69, 128]. In contrast to conventional spatial adapter modules, the temporal adapter module employs depth-wise 3D convolution instead of an intermediate activation layer [36]. The temporal adapter could be formally written as:

$$\texttt{T-Adapter}(\mathbf{X}) = \mathbf{X} + \mathbf{W}_{\texttt{up}}(\texttt{3D-Conv}(\mathbf{W}_{\texttt{down}}(\mathbf{X}))). \tag{5}$$

By utilizing 3D convolutions in lower-dimensional input, our approach significantly alleviates the complexity of temporal modeling. As a result, our method achieves efficient memory usage during training and exhibits the fastest inference speed among competitive approaches, as demonstrated in Table 1.

Temporal Latent-Shift Attention In the original T2I framework, the attention block of the U-Net only performs self-attention for individual frames, neglecting the information from other frames. While joint-space-time attention, as demonstrated in [2, 7, 96, 114], can effectively model temporal dependencies, it introduces a quadratic complexity in terms of attention calculation. For a video with L frames and N tokens, the complexity of global spatial-temporal attention becomes $O(L^2N^2)$. To address this issue, we propose a latent-shift attention module as shown in Fig. 3. In addition to considering tokens within the current frame, we further conduct a patch-level shifting operation along the temporal dimension to shift tokens from the preceding T frames onto the current frame, thereby composing a new latent feature frame. We concatenate the latent feature of the current frame with the temporally shifted latent feature, forming the keys and values for attention calculation. The latent-shift attention can be formally written as:

$$\mathbf{Q} = \mathbf{W}_{\texttt{q}}(\boldsymbol{x}_{z_i}), \tag{6}$$

$$\mathbf{K} = \mathbf{W}_{\texttt{k}}[\boldsymbol{x}_{z_i}, \boldsymbol{x}_{z_{shift}}], \tag{7}$$

$$\mathbf{V} = \mathbf{W}_{\texttt{v}}[\boldsymbol{x}_{z_i}, \boldsymbol{x}_{z_{shift}}], \tag{8}$$

where $\boldsymbol{x}_{z_i}$ denotes the query frame and $[\cdot]$ means concatenate. This approach reduces the complexity of attention to $O(2LN^2)$, significantly lowering the computational burden compared to global attention. Moreover, it allows the model to learn the relationships between adjacent frames, ensuring better temporal consistency in video generation.

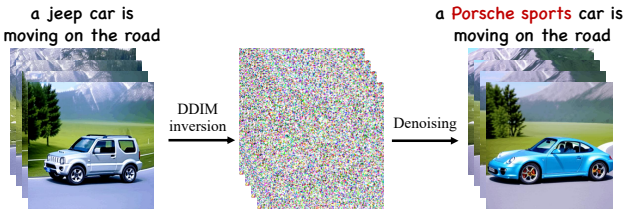


Figure 4. During inference, we sample a novel video from the latent noise inverted from the input video, guided by an edited prompt (*e.g.*, "a Porsche sports car is moving on the road").

3.4. Super Resolution and Editing Models

Super Resolution (SR) Due to constraints of limited GPU memory and the lack of high-resolution video-text datasets, most existing methods [1, 8, 35], including ours, are only able to generate images at a resolution of 256×256. To overcome this limitation and generate higher-resolution outputs, we adopt a two-stage training approach similar to cascaded Diffusion Models [40]. In the first stage, we generate videos with a 256×256 resolution using our SimDA methods. In the second stage, we employ an LDM upsampler [79] to enhance the resolution of the videos to 1024×1024. We incorporate noise augmentation and noise level conditioning, and train a super-resolution model using the following equation:

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta([\boldsymbol{x}_t, \boldsymbol{x}_{low}], \boldsymbol{c}, t)||_2^2], \tag{9}$$

where $\boldsymbol{x}_{low}$ is the low-resolution video, we concatenate it with $\boldsymbol{x}_t$ frame by frame following Video LDM [8]. The architecture of SR is similar to T2V model in the first stage,

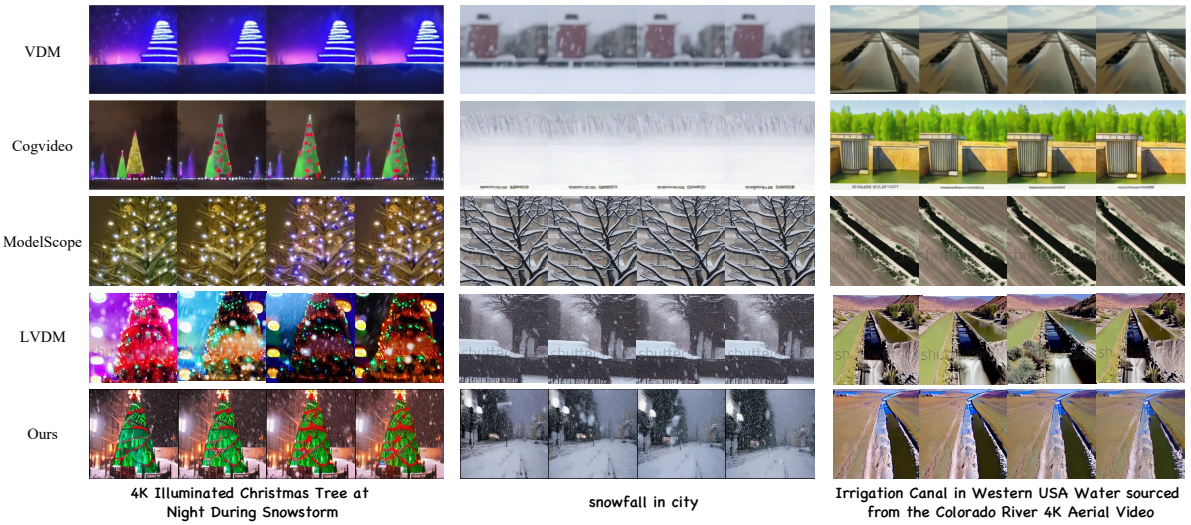


Figure 5. Text-to-Video generation comparison on the user study evaluation set.

Table 2. Text-to-Video generation comparison on MSR-VTT [118] dataset. We report the Fréchet Video Distance (FVD) scores and CLIPSIM scores.

Method	Training Data	Resolution	Zero-shot	Params(B)	FVD(↓)	CLIPSIM(↑)
GODIVA [108]	MSR-VTT	128x128	No	-	-	0.2402
NÜWA [109]	MSR-VTT	128x128	No	0.87	-	0.2439
Make-A-Video [88]	WebVid-10M + HD-VILA-10M	256x256	Yes	9.72	-	0.3049
VideoFactory [104]	WebVid-10M + HD-VG-130M	256x256	Yes	2.04	-	0.3005
LVDM [35]	WebVid-2M	256x256	Yes	1.16	742	0.2381
MMVG [25]	WebVid-2.5M	256x256	Yes	-	-	0.2644
CogVideo [42]	WebVid-5.4M	256x256	Yes	15.5	1294	0.2631
ED-T2V [57]	WebVid-10M	256x256	Yes	1.30	-	0.2763
MagicVideo [136]	WebVid-10M	256x256	Yes	-	998	-
Video-LDM [8]	WebVid-10M	256x256	Yes	4.20	-	0.2929
VideoComposer [105]	WebVid-10M	256x256	Yes	1.85	580	0.2932
Latent-Shift [1]	WebVid-10M	256x256	Yes	1.53	-	0.2773
VideoFusion [61]	WebVid-10M	256x256	Yes	1.83	581	0.2795
SimDA (Ours)	WebVid-10M	256x256	Yes	**1.08**	**456**	**0.2945**

we change the original U-Net block by adding Spatial and Temporal Adapters as described in Sec. 3.3 and only fine-tune the newly added modules.

Text-guided Video Editing In addition to performing T2V generation, our method could turn into one-shot tuning for text-guided video editing following Tune-A-Video [110]. The training pipeline of editing model is the same as our T2V method. However, for the inference stage, we adopt the DDIM inversion latents instead of random noisy latents together with edited prompt for novel video generation as shown in Fig. 4. By doing so, the pixel-level information control could remain in the inversion latent as demonstrated in [110]. Owing to the light-weight module and efficient pipeline of our method, SimDA needs fewer training steps (200 steps compared to 500 steps) and thus the training time and inference time is much faster than Tune-A-Video [110].

4. Experiments

4.1. Implementation Details

Our T2V method is composed of two-stage models. The first model predicts video frames with a resolution 256×256 (with a latent size of 32×32), while the second model is a $4\times$ upsampler, producing a resolution of 1024×1024. We train the general T2V model on WebVid-10M [5] dataset following [1, 8]. We follow previous methods [1, 104, 136] to report the CLIP score [75] and FVD (Fréchet Video Distance) score [99] on MSR-VTT [118]. Besides, we compare the FVD score and CLIP score on the evaluation set of WebVid [5] as in VideoFactory [104]. We also compare the parameter scale and inference speed of our method with some open-sourced methods [1, 35, 61]. Finally, we also provide a user study between our work and VDM [41], Latent-shift [1], Video-Fusion [61] and LVDM [35].

4.2. Evaluation on Text-to-Video Generation

To fully evaluate the generation performance of our SimDA, we conduct automatic evaluations on two distinct datasets: WebVid [5] (Val), which shares the same domain as the training data and MSR-VTT [118] in a zero-shot setting.

Evaluation on MSR-VTT As shown in Table 2, we evaluate CLIPSIM [75] and FVD [99] on the widely used video generation benchmarks, MSR-VTT [118]. We randomly select one text prompt per example from MSR-VTT [118] and generate a total of 2,990 videos. Despite being a zero-shot setting, our method achieves an average CLIPSIM of 0.2945 that surpasses most of the competitors, indicating a strong semantic alignment between the generated videos and the input text. Though Make-A-Video [88] and VideoFactory [104] offer higher CLIP scores, they utilize additional large-scale HD-VILA [119] datasets for training.

Evaluation on WebVid As shown in Table 3, we create a validation set consisting of 4,476 randomly extracted text-video pairs from WebVid-10M. These pairs are not included in the training data following [104]. We conduct evaluations on this validation set and obtain impressive results. Our method achieves an FVD score of 363.98 and a CLIPSIM score of 0.3054. These scores are significantly higher than those achieved by existing methods such as ModelScope [103] and LVDM [35]. Besides, our method shows competitive results compared to VideoFactory [104] which is trained with much larger datasets. These results clearly demonstrate the superiority of our approach.

Human Evaluation In order to address the limitations of existing metrics and assess the performance of our SimDA from a human perspective, we conduct an extensive user study. The study involves comparing our method with four state-of-the-art methods. Specifically, we select two publicly available models, namely ModelScope [103] and LVDM [35]. Additionally, we consider two methods with similar scale parameters, VDM [41] and Latent-shift [1], which only showcase some samples on their websites.

For each case, participants were provided with two video samples, one is generated by our method and the other is from a competitor. They were then asked to compare the two samples in terms of video quality and text-video similarity. To ensure fairness in the comparisons, we also report the ratio of network parameters compared to ours. The results, along with the parameter ratios, are presented in Table 4. The user study approach allows us to gain in-depth insights into the subjective evaluation of our method.

Qualitative Results The visualization of T2V generation results are shown in Fig. 1(a). Besides, we show the comparison results in Fig. 5. More examples can be found in our supplementary material.

Parameter Size and Inference Speed We conduct a comparison of the number of parameters and inference speed, and the results are presented in Table 1. For the speed comparison, we select CogVideo [42], Latent-Shift [1] and LVDM [35]. SimDA, on the other hand, stands out as it is significantly smaller than previous works and exhibits faster inference speed compared to other methods. Despite having fewer parameters, SimDA achieves superior performance in various benchmarks when compared to other methods. This validation further highlights our advantages in terms of model efficiency and performance.

Table 3. Text-to-video generation on the validation set of WebVid [5]. We report the FVD and CLIPSIM scores.

Method	Params(B)	FVD(↓)	CLIPSIM(↑)
LVDM [35]	1.16	455.53	0.2751
ModelScope [103]	1.83	414.11	0.3000
VideoFactory [104]	2.04	292.35	0.3070
SimDA (Ours)	**1.08**	**363.98**	**0.3054**

Table 4. User preference is depicted as a percentage indicating the proportion of individuals favoring our method over the compared approach. Param Ratio means the ratio of the network parameter *v.s.* Ours.

Sample	Method	Param Ratio	Quality	Faithfulness
Open Website	VDM [41]	0.83×	85.2%	81.4%
	Latent-Shift [1]	1.41×	81.5%	79.3%
Pretrained Model	ModelScope [103]	1.69×	78.3%	79.5%
	LVDM [35]	1.07×	83.4%	84.7%

4.3. Evaluation on Text-guided Video Editing

Following the methodology of previous studies [110], we employ CLIP score [75] and a user study to evaluate the performance of different methods in terms of frame consistency and textual alignment.

First, we calculate the CLIP image embedding for all frames in the edited videos to measure frame consistency. The average cosine similarity between pairs of video frames is reported. Additionally, to assess textual faithfulness, we compute the average CLIP score between frames of the output videos and the corresponding edited prompts. A total of 15 videos from the dataset [72] were selected and edited based on object, background and style, resulting in 75 edited videos for each model. The average results, presented in Table 5, highlight our method's exceptional ability to achieve semantic alignment.

Secondly, we conduct a user study involving videos and text prompts. Participants were asked to vote for the edited videos that exhibited the best temporal consistency and those most accurately matched the textual description. Table 5 demonstrates that our method, SimDA, receives the highest number of votes in both aspects, indicating superior editing quality and a strong preference from users in practical scenarios.

Table 5. Quantitative comparison with evaluated baseline [110]. The "Training" refers to the process of optimization, and "Memory" refers to the GPU memory.

Method	Frame consistency		Textual alignment		Runtime [min]		Memory [Gib]		Params [Mb]
	CLIP Score↑	User Vote↑	CLIP Score↑	User Vote↑	Training↓	Inference↓	Training↓	Inference↓	Tuned↓
Tune-A-Video [110]	94.1	31.2%	31.8	39.5%	9.3	0.8	31.3	11.4	74.4
SimDA(Ours)	**94.9**	**68.8%**	**31.9**	**60.5%**	**2.5**	**0.4**	**28.6**	**8.8**	**24.9**

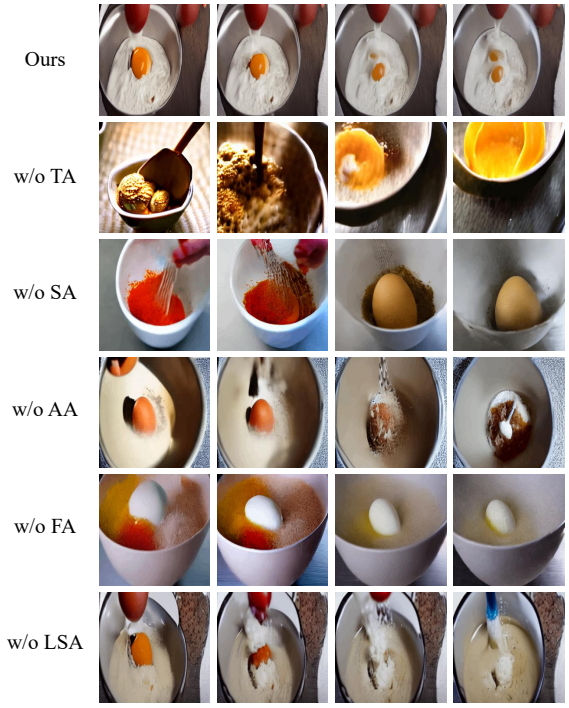


Figure 6. Ablation of T2V generation results. TA, SA, AA, FA and LSA refer to Temporal Adapter, Spatial Adapter, Attention Adapter, FFN Adapter and Latent-Shift Attention.

4.4. Ablation Study

Here we discuss the effect of each module. We perform experiments on $1K$ samples from the WebVid validation set.

Effect of Temporal Adapter Temporal modeling is a crucial component of video generation. In our video editing task, when compared to methods that rely on temporal attention modeling like Tune-A-Video [110], we observe that our temporal adapter is more lightweight and achieves superior editing results as in Table 5. Additionally, we conduct ablation experiments, as shown in Table 6 and Fig. 6, where the lack of Temporal Adapter (TA) results in significantly higher FVD and chaotic temporal sequences in the generated videos.

Effect of Spatial Adapter We also validate the effectiveness of the Spatial Adapter (SA) in transferring spatial knowledge of videos. As shown in Table 6, without the Attention Adapter (AA) and FFN Adapter (FA), the model's FVD and CLIPSIM scores for generated videos will become worse. Additionally, it can be observed from the Fig. 6 that the model exhibits misconceptions in understanding the text prompt without the spatial adapter.

Table 6. Ablation study on different modules. We report the FVD [99] and CLIPSIM [75] on $1K$ samples from the validation set of WebVid-10M [5]. TA, SA, AA, FA and LSA represent Temporal Adapter, Spatial Adapter, Attention Adapter, FFN Adapter and Latent-shift Attention, respectively.

	TA	AA	FA	LSA	FVD(↓)	CLIPSIM(↑)
w/o TA		✓	✓	✓	1470.1	0.2629
w/o SA	✓			✓	811.3	0.2822
w/o AA	✓		✓	✓	764.8	0.2851
w/o FA	✓	✓		✓	623.7	0.2962
w/o LSA	✓	✓	✓		618.2	0.3011
Ours	✓	✓	✓	✓	**530.2**	**0.3034**

Effect of Latent Shift Attention To investigate the impact of Latent-shift Attention (LSA), we replace it with regular single-frame spatial attention. Besides observing a decline in FVD and text alignment CLIPSIM scores in Table 6, we also test the CLIPSIM of each frame within the same video, which decreased from 96.4 to 94.5. This demonstrates that our LSA module can effectively model the relationship of adjacent frames, leading to more consistent videos.

5. Conclusion

In this paper, we proposed SimDA, a parameter-efficient video diffusion model for text-guided video generation and editing. With the proposed light-weight spatial and temporal adapters, our method not only transferred from powerful spatial information but also modeled temporal relationships with the least new parameters. The experimental results demonstrated that our approach has the fastest training and inference speed while maintaining competitive generation and editing results. Our work is the first parameter-efficient video diffusion method serving as an efficient T2V fine-tuning baseline and paved the way for future research.

Acknowledge This work was supported by National Science and Technology Major Project (No. 2021ZD0112805) and in part by National Natural Science Foundation of China (No. 62032006 and No. 62102092.). The computations in this research were performed using the CFFFplatform of Fudan University.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 2, 3, 4, 5, 6, 7

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 5

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017. 2

[4] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 6, 7, 8

[6] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *ECCV*, 2022. 3

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 5

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2, 4, 5, 6

[9] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *NeurIPS*, 2022. 2

[10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 5

[12] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *ICCV*, 2019. 2

[13] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *NeurIPS*, 2022. 2, 3, 4

[14] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 2

[15] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018. 2

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2

[18] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *NeurIPS*, 2022. 2

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[20] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[21] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3

[22] Qijun Feng, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Fdgaussian: Fast gaussian splatting from single image via geometric-aware diffusion model. *arXiv preprint arXiv:2403.10242*, 2024. 3

[23] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*, 2023. 2

[24] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *ICML*, 2020. 2

[25] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *CVPR*, 2023. 6

[26] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022. 3

[27] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 2

[28] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *arXiv preprint arXiv:2305.10474*, 2023. 2

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 2

[30] Sonam Gupta, Arti Keshari, and Sukhendu Das. Rv-gan: Recurrent gan for unconditional video generation. In *CVPR*, 2022. 2

[31] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, 2018. 2

[32] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. 2

[33] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 3

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[35] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 4, 5, 6, 7

[36] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[37] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ICLR*, 2023. 3, 4

[38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[39] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 4

[40] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 2, 5

[41] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 6, 7

[42] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 4, 6, 7

[43] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 2

[44] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019. 2, 3

[45] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3, 4

[46] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 3

[47] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 3

[48] Emmanuel Kahembwe and Subramanian Ramamoorthy. Lower dimensional kernels for video discriminators. *Neural Networks*, 132:506–520, 2020. 2

[49] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[50] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[51] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 2021. 2

[52] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 2

[53] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2, 3

[54] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2, 3, 4

[55] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018. 2

[56] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 3

[57] Jiawei Liu, Weining Wang, Wei Liu, Qian He, and Jing Liu. Ed-t2v: An efficient training framework for diffusion-based text-to-video generation. In *IJCNN*, 2023. 6

[58] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2

[59] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3

[60] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020. 2

[61] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 2, 6

[62] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *ICCV*, 2017. 2

[63] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *ICLR*, 2022. 3

[64] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using

deep recurrent attentive architectures. In *ACM Multimedia*, 2017. 2

[65] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 4

[66] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift gan for large scale video generation. In *WACV*, 2021. 3

[67] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[68] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[69] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. *NeurIPS*, 2022. 2, 3, 5

[70] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *ACM Multimedia*, 2017. 2

[71] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In *CVPR*, 2024. 2

[72] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 7

[73] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, 2023. 3

[74] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6, 7, 8

[76] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020. 2

[77] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

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

[79] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5

[80] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[81] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 4

[82] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2

[83] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *IJCV*, 2020. 2

[84] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022. 2

[85] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *CVPR*, 2020. 5

[86] Baifeng Shi, Qi Dai, Judy Hoffman, Kate Saenko, Trevor Darrell, and Huijuan Xu. Temporal action detection with multi-level supervision. In *ICCV*, 2021. 5

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

[88] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023. 2, 4, 5, 6, 7

[89] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 2

[90] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[91] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[92] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[93] Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *NeurIPS*, 2021. 3

[94] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021. 2

[95] Shuyuan Tu, Tianzhen Guan, and Li Kuang. Multiple biological granularities network for person re-identification. In *ICMR*, 2022. 3

[96] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *ICCV*, 2023. 5

[97] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024. 3

[98] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 3

[99] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6, 7, 8

[100] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 2

[101] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*, 2022. 2

[102] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *NeurIPS*, 2016. 2

[103] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 7

[104] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 2, 5, 6, 7

[105] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 6

[106] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, Kai Qiu, Yuhui Yuan, Chuanxin Tang, Xiaoyan Sun, Chong Luo, and Baining Guo. Microcinema: A divide-and-conquer approach for text-to-video generation. In *CVPR*, 2024. 2

[107] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 2

[108] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2, 6

[109] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 6

[110] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 2, 3, 6, 7, 8

[111] Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Xian-Sheng Hua, and Lei Zhang. Spatiotemporal self-attention modeling with temporal patch shift for action recognition. In *ECCV*, 2022. 3

[112] Zhen Xing, Yijiang Chen, Zhixin Ling, Xiangdong Zhou, and Yu Xiang. Few-shot single-view 3d reconstruction with memory prior contrastive network. In *ECCV*, 2022. 3

[113] Zhen Xing, Hengduo Li, Zuxuan Wu, and Yu-Gang Jiang. Semi-supervised single-view 3d reconstruction via prototype shape priors. In *ECCV*, 2022. 3

[114] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *CVPR*, 2023. 5

[115] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023. 3

[116] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023. 2

[117] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *CVPR*, 2018. 2

[118] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 6, 7

[119] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 7

[120] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv preprint arXiv:2305.18295*, 2023. 2

[121] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[122] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 2

[123] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 3

[124] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *ICLR*, 2023. 2, 3, 4

[125] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku,

Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[126] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571*, 2022. 2

[127] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 3

[128] Bowen Zhang, Xiaojie Jin, Weibo Gong, Kai Xu, Zhao Zhang, Peng Wang, Xiaohui Shen, and Jiashi Feng. Multimodal video adapter for parameter efficient video text retrieval. *arXiv preprint arXiv:2301.07868*, 2023. 2, 3, 5

[129] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *ACM Multimedia*, 2021. 3

[130] Hui Zhang, Zuxuan Wu, Zhen Xing, Jie Shao, and Yu-Gang Jiang. Adadiff: Adaptive step selection for fast diffusion. *arXiv preprint arXiv:2311.14768*, 2023. 2

[131] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4

[132] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[133] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *ICLR*, 2023. 3

[134] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. *arXiv preprint arXiv:2305.17431*, 2023. 3

[135] Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In *CVPR*, 2024. 2

[136] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 6

[137] Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer. In *CVPR*, 2024. 2