# DreamVideo: Composing Your Dream Videos with Customized Subject and Motion

Yujie Wei[1] Shiwei Zhang[2] Zhiwu Qing[2] Hangjie Yuan[2] Zhiheng Liu[2]
Yu Liu[2] Yingya Zhang[2] Jingren Zhou[2] Hongming Shan[1,3,4†]
[1] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University
[2] Alibaba Group [3] MOE Frontiers Center for Brain Science, Fudan University
[4] Key Laboratory of Computational Neuroscience and Brain-inspired Intelligence, Fudan University
yjwei22@m.fudan.edu.cn, hmshan@fudan.edu.cn, {zhangjin.zsw, qingzhiwu.qzw}@alibaba-inc.com
{yuanhangjie.yhj, pingzhi.lzh, ly103369, yingya.zyy, jingren.zhou}@alibaba-inc.com

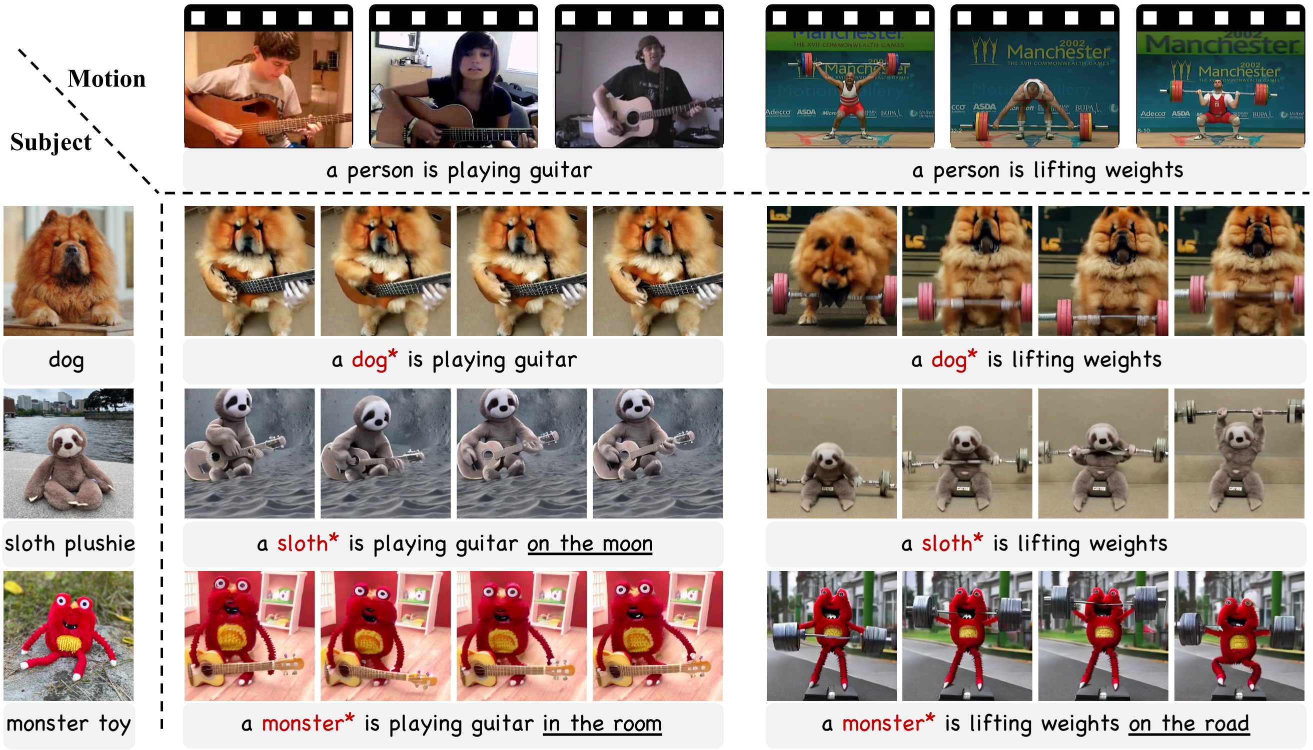

Motion
Subject
a person is playing guitar
a person is lifting weights
dog
a dog* is playing guitar
a dog* is lifting weights
sloth plushie
a sloth* is playing guitar on the moon
a sloth* is lifting weights
monster toy
a monster* is playing guitar in the room
a monster* is lifting weights on the road

Figure 1. **Customized video generation results** of our proposed DreamVideo with specific subjects (left) and motions (top). Our method can customize *both* subject identity and motion pattern to generate desired videos with various context descriptions.

## Abstract

*Customized generation using diffusion models has made impressive progress in image generation, but remains unsatisfactory in the challenging video generation task, as it requires the controllability of both subjects and motions. To that end, we present DreamVideo, a novel approach to generating personalized videos from a few static images of the desired subject and a few videos of target motion. DreamVideo decouples this task into two stages, subject learning and motion learning, by leveraging a pre-trained video diffusion model. The subject learning aims to accurately capture the fine appearance of the subject from provided images, which is achieved by combining textual inversion and fine-tuning of our carefully designed identity adapter. In motion learning, we architect a motion adapter*

† Corresponding author.

*and fine-tune it on the given videos to effectively model the target motion pattern. Combining these two lightweight and efficient adapters allows for flexible customization of any subject with any motion. Extensive experimental results demonstrate the superior performance of our DreamVideo over the state-of-the-art methods for customized video generation. Our project page is at `https://dreamvideo-t2v.github.io`.*

## 1. Introduction

The remarkable advances in diffusion models [5, 29, 46, 56, 64] have empowered designers to generate photorealistic images and videos based on textual prompts, paving the way for customized content generation [18, 57]. While customized image generation has witnessed impressive progress [2, 38, 44, 45], the exploration of customized video generation remains relatively limited. The main reason is that videos have diverse spatial content and intricate temporal dynamics simultaneously, presenting a highly challenging task to concurrently customize these two key factors.

Current existing methods [50, 77] have effectively propelled progress in this field, but they are still limited to optimizing a single aspect of videos, namely spatial subject or temporal motion. For example, Dreamix [50] and Tune-A-Video [77] optimize the spatial parameters and spatial-temporal attention to inject a subject identity and a target motion, respectively. However, focusing only on one aspect (*i.e.*, subject or motion) may reduce the model's generalization on the other aspect. On the other hand, AnimateDiff [23] trains temporal modules appended to the personalized text-to-image models for image animation. It tends to pursue generalized video generation but suffers from a lack of motion diversity, such as focusing more on camera movements, making it unable to meet the requirements of customized video generation tasks well. Therefore, we believe that effectively modeling both spatial subject and temporal motion is necessary to enhance video customization.

The above observations drive us to propose the DreamVideo, which can synthesize videos featuring the user-specified subject endowed with the desired motion from a few images and videos respectively, as shown in Fig. 1. DreamVideo decouples video customization into subject learning and motion learning, which can reduce model optimization complexity and increase customization flexibility. In subject learning, we initially optimize a textual identity using Textual Inversion [18] to represent the coarse concept, and then train a carefully designed identity adapter with the frozen textual identity to capture fine appearance details from the provided static images. In motion learning, we design a motion adapter and train it on the given videos to capture the inherent motion pattern. To avoid the shortcut of learning appearance features at this stage, we incorporate the image feature into the motion adapter to enable it to concentrate exclusively on motion learning. Benefiting from these two-stage learning, DreamVideo can flexibly compose customized videos with any subject and any motion once the two lightweight adapters have been trained.

To validate DreamVideo, we collect 20 customized subjects and 30 motion patterns as a substantial experimental set. The extensive experimental results unequivocally showcase its remarkable customization capabilities surpassing the state-of-the-art methods.

In summary, our main contributions are:

1. We propose DreamVideo, a novel approach for customized video generation with any subject and motion. To the best of our knowledge, this work makes the first attempt to customize *both* subject identity and motion.
2. We propose to decouple the learning of subjects and motions by the devised identity and motion adapters, which can greatly improve the flexibility of customization.
3. We conduct extensive qualitative and quantitative experiments, demonstrating the superiority of DreamVideo over the existing state-of-the-art methods.

## 2. Related Work

**Text-to-video generation.** Text-to-video generation aims to generate realistic videos based on text prompts and has received growing attention [9, 10, 14, 16, 25, 32, 37, 40, 47, 49, 82, 85]. Early works are mainly based on Generative Adversarial Networks (GANs) [4, 36, 59, 62, 67–69] or autoregressive transformers [19, 33, 39, 81]. Recently, to generate high-quality and diverse videos, many works apply the diffusion model to video generation [1, 6, 17, 20, 21, 26, 35, 43, 53, 54, 71, 73–75, 78, 86, 88, 89]. Make-A-Video [61] leverages the prior of the image diffusion model to generate videos without paired text-video data. Video Diffusion Models [31] and ImagenVideo [30] model the video distribution in pixel space by jointly training from image and video data. To reduce the huge computational cost, VLDM [7] and MagicVideo [91] apply the diffusion process in the latent space, following the paradigm of LDMs [56]. ModelScopeT2V [70] and VideoComposer [72] incorporate spatiotemporal blocks with various conditions for controllable video generation. These powerful video generation models pave the way for customized video generation.

**Customized generation.** Compared with general generation tasks, customized generation may better accommodate user preferences. Most current works focus on subject customization with a few images [11, 13, 24, 58, 60, 63, 76]. Textual Inversion [18] represents a user-provided subject through a learnable text embedding without model fine-tuning. DreamBooth [57] tries to bind a rare word with a subject by fully fine-tuning an image diffusion model. Moreover, some works study the more challenging multi-subject customization task [15, 22, 38, 44, 45, 48, 79]. De-

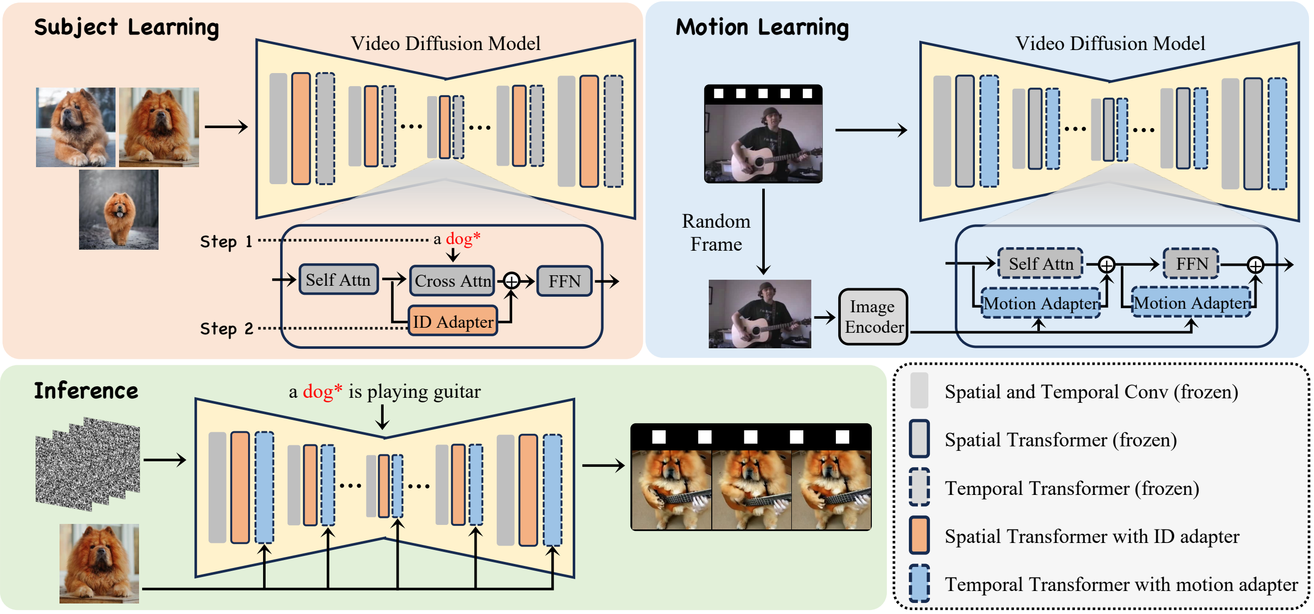


Figure 2. **Illustration of the proposed DreamVideo**, which decouples customized video generation into two stages. In subject learning, we first optimize a unique textual identity for the subject, and then train the devised identity adapter (ID adapter) with the frozen textual identity to capture fine appearance details. In motion learning, we pass a randomly selected frame from its training video through the CLIP image encoder, and use its embedding as the appearance condition to guide the training of the designed motion adapter. Note that we freeze the pre-trained video diffusion model throughout the training process. During inference, we combine the two lightweight adapters and randomly select an image provided during training as the appearance guidance to generate customized videos.

spite the significant progress in customized image generation, customized video generation is still under exploration. Dreamix [50] attempts subject-driven video generation by following the paradigm of DreamBooth. However, fine-tuning the video diffusion model tends to overfitting and generate videos with small or missing motions. A concurrent work [90] aims to customize the motion from training videos. Nevertheless, it fails to customize the subject, which may be limiting in practical applications. In contrast, this work proposes DreamVideo to effectively generate customized videos with *both* specific subject and motion.

**Parameter-efficient fine-tuning.** Drawing inspiration from the success of parameter-efficient fine-tuning (PEFT) in NLP [34, 41] and vision tasks [3, 12, 83, 84], some works adopt PEFT for video generation and editing tasks due to its efficiency [51, 80]. In this work, we explore the potential of lightweight adapters, revealing their superior suitability for customized video generation.

## 3. Methodology

In this section, we first introduce the preliminaries of Video Diffusion Models. We then present DreamVideo to showcase how it can compose videos with the customized subject and motion. Finally, we analyze the efficient parameters for subject and motion learning while describing training and inference processes for our DreamVideo.

### 3.1. Preliminary: Video Diffusion Models

Video diffusion models (VDMs) [7, 31, 70, 72] are designed for video generation tasks by extending the image diffusion models [29, 56] to adapt to the video data. VDMs learn a video data distribution by the gradual denoising of a variable sampled from a Gaussian distribution. This process simulates the reverse process of a fixed-length Markov Chain. Specifically, the diffusion model $\epsilon_\theta$ aims to predict the added noise $\epsilon$ at each timestep $t$ based on text condition $c$, where $t \in \mathcal{U}(0,1)$. The training objective can be simplified as a reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{z,c,\epsilon\sim\mathcal{N}(0,\mathrm{I}),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, \tau_\theta(c), t\right)\right\|_2^2\right], \tag{1}$$

where $z \in \mathbb{R}^{B\times F\times H\times W\times C}$ is the latent code of video data with $B, F, H, W, C$ being batch size, frame, height, width, and channel, respectively. $\tau_\theta$ presents a pre-trained text encoder. A noise-corrupted latent code $z_t$ from the ground-truth $z_0$ is formulated as $z_t = \alpha_t z_0 + \sigma_t \epsilon$, where $\sigma_t = \sqrt{1-\alpha_t^2}$, $\alpha_t$ and $\sigma_t$ are hyperparameters to control the diffusion process. Following ModelScopeT2V [70], we instantiate $\epsilon_\theta(\cdot,\cdot,t)$ as a 3D UNet, where each layer includes a spatiotemporal convolution layer, a spatial transformer, and a temporal transformer, as shown in Fig. 2.

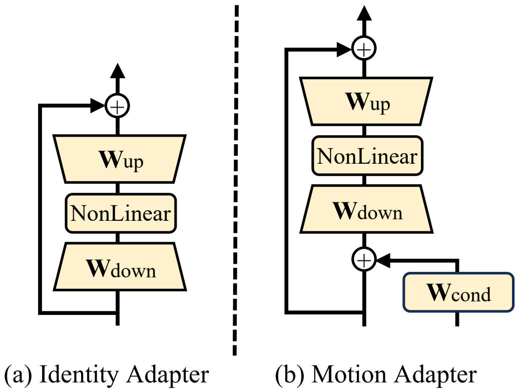


Figure 3. **Illustration of the devised adapters.** Both use a bottleneck structure. Compared to identity adapter, motion adapter adds a linear layer to incorporate the appearance guidance.

### 3.2. DreamVideo

Given a few images of one subject and multiple videos (or a single video) of one motion pattern, our goal is to generate customized videos featuring both the specific subject and motion. To this end, we propose DreamVideo, which decouples the challenging customized video generation task into subject learning and motion learning via two devised adapters, as illustrated in Fig. 2. Users can simply combine these two adapters to generate desired videos.

**Subject learning.** To accurately preserve subject identity and mitigate overfitting, we introduce a two-step training strategy inspired by [2] for subject learning with 3∼5 images, as illustrated in the upper left portion of Fig. 2.

The first step is to learn a textual identity using Textual Inversion [18]. We freeze the video diffusion model and only optimize the text embedding of pseudo-word "$S^*$" using Eq. (1). The textual identity represents the coarse concept and serves as a good initialization.

Leveraging only the textual identity is not enough to reconstruct the appearance details of the subject, so further optimization is required. Instead of fine-tuning the video diffusion model, our second step is to learn a lightweight identity adapter by incorporating the learned textual identity. We freeze the text embedding and only optimize the parameters of the identity adapter. As demonstrated in Fig. 3(a), the identity adapter adopts a bottleneck architecture with a skip connection, which consists of a down-projected linear layer with weight $\mathbf{W}_{\text{down}} \in \mathbb{R}^{l\times d}$, a nonlinear activation function $\sigma$, and an up-projected linear layer with weight $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d\times l}$, where $l > d$. The adapter training process for the input spatial hidden state $h_t \in \mathbb{R}^{B\times(F\times h\times w)\times l}$ can be formulated as:

$$h'_t = h_t + \sigma\left(h_t * \mathbf{W}_{\text{down}}\right) * \mathbf{W}_{\text{up}}, \tag{2}$$

where $h, w, l$ are height, width, and channel of the hidden feature map, $h'_t$ is the output of identity adapter, and $F = 1$ because only image data is used. We employ GELU [27] as the activation function $\sigma$. In addition, we initialize $\mathbf{W}_{\text{up}}$ with zeros to protect the pre-trained diffusion model from being damaged at the beginning of training [87].

**Motion learning.** Another important property of customized video generation is to make the learned subject move according to the desired motion pattern from existing videos. To efficiently model a motion, we devise a motion adapter with a structure similar to the identity adapter, as depicted in Fig. 3(b). Our motion adapter can be customized using a motion pattern derived from a class of videos (*e.g.*, videos representing various dog motions), multiple videos exhibiting the same motion, or even a single video.

Although the motion adapter enables capturing the motion pattern, it inevitably learns the appearance of subjects from the input videos during training. To disentangle spatial and temporal information, we incorporate appearance guidance into the motion adapter, forcing it to learn pure motion. Specifically, we add a condition linear layer with weight $\mathbf{W}_{\text{cond}} \in \mathbb{R}^{C'\times l}$ to integrate appearance information into the temporal hidden state $\hat{h}_t \in \mathbb{R}^{(B\times h\times w)\times F\times l}$. Then, we randomly select one frame from the training video and pass it through the CLIP [55] image encoder to obtain its image embedding $e \in \mathbb{R}^{B\times 1\times C'}$. This image embedding is subsequently broadcasted across all frames, serving as the appearance guidance during training. The forward process of the motion adapter is formulated as:

$$\hat{h}^e_t = \hat{h}_t + \text{broadcast}(e * \mathbf{W}_{\text{cond}}), \tag{3}$$

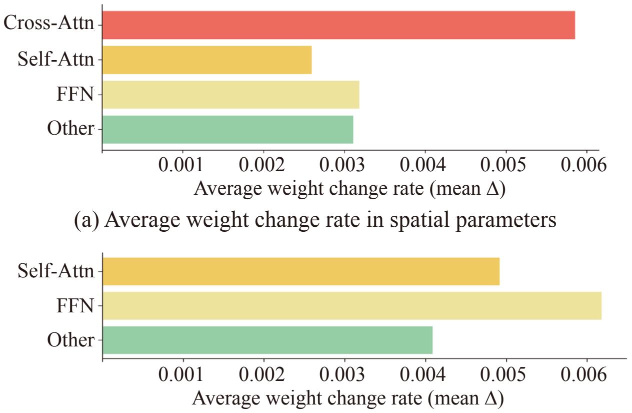
$$\hat{h}'_t = \hat{h}_t + \sigma(\hat{h}^e_t * \mathbf{W}_{\text{down}}) * \mathbf{W}_{\text{up}}, \tag{4}$$

where $\hat{h}'_t$ is the output of motion adapter. At inference time, we randomly take a training image provided by the user as the appearance condition input to the motion adapter.

### 3.3. Model Analysis, Training and Inference

**Where to put these two adapters.** We address this question by analyzing the change of all parameters within the fine-tuned model to determine the appropriate position of the adapters. These parameters are divided into four categories: (1) cross-attention (only exists in spatial parameters), (2) self-attention, (3) feed-forward, and (4) other remaining parameters. Following [38, 42], we use $\Delta_l = \|\theta'_l - \theta_l\|_2 \,/\, \|\theta_l\|_2$ to calculate the weight change rate of each layer, where $\theta'_l$ and $\theta_l$ are the updated and pre-trained model parameters of layer $l$. Specifically, to compute $\Delta$ of spatial parameters, we only fine-tune the spatial parameters of the UNet while freezing temporal parameters, for which the $\Delta$ of temporal parameters is computed in a similar way.

We observe that the conclusions are different for spatial and temporal parameters. Fig. 4(a) shows the mean $\Delta$ of spatial parameters for the four categories when fine-tuning



(a) Average weight change rate in spatial parameters

(b) Average weight change rate in temporal parameters

Figure 4. **Analysis of weight change** on updating all spatial or temporal model weights during fine-tuning. We observe that cross-attention layers play a key role in subject learning while the contributions of all layers are similar to motion learning.

the model on "Chow Chow" images (dog in Fig. 1). The result suggests that the cross-attention layers play a crucial role in learning appearance compared to other parameters. However, when learning motion dynamics in the "bear walking" video (see Fig. 7), all parameters contribute close to importance, as shown in Fig. 4(b). Remarkably, our findings remain consistent across various images and videos. This phenomenon reveals the divergence of efficient parameters for learning subjects and motions. Therefore, we insert the identity adapter to cross-attention layers while employing the motion adapter to all layers in temporal transformer.

**Decoupled training strategy.** Customizing the subject and motion simultaneously on images and videos requires training a separate model for each combination, which is time-consuming and impractical for applications. Instead, we tend to decouple the training of subject and motion by optimizing the identity and motion adapters independently according to Eq. (1) with the frozen pre-trained model.

**Inference.** During inference, we combine the two customized adapters and randomly select an image provided during training as the appearance guidance to generate customized videos. We find that choosing different images has a marginal impact on generated results. Besides combinations, users can also customize the subject or motion individually using only the identity adapter or motion adapter.

## 4. Experiment

### 4.1. Experimental Setup

**Datasets.** For subject customization, we select subjects from image customization papers [22, 45, 57] for a total of 20 subjects, including 9 pets and 11 objects. For motion customization, we collect a dataset of 30 motion patterns from the UCF101 dataset [66], the UCF Sports Action dataset [65], and the DAVIS dataset [52]. We also provide 42 text prompts used for extensive experimental validation, where the prompts are designed to generate new motions of subjects, new contexts of subjects and motions, and *etc*.

**Implementation details.** For subject learning, we take $\sim$3000 iterations for optimizing the textual identity following [18, 45] with learning rate $1.0 \times 10^{-4}$, and $\sim$800 iterations for learning identity adapter with learning rate $1.0 \times 10^{-5}$. For motion learning, we train motion adapter for $\sim$1000 iterations with learning rate $1.0 \times 10^{-5}$. During inference, we employ 50-step DDIM [64] and classifier-free guidance [28] to generate 32-frame videos with 8 fps.

**Baselines.** Since there is no existing work for customizing both subjects and motions, we consider comparing our method with three categories of combination methods: AnimateDiff [23], ModelScopeT2V [70], and LoRA fine-tuning [34]. AnimateDiff trains a motion module appended to a pre-trained image diffusion model from Dreambooth [57]. However, we find that training from scratch leads to unstable results. For a fair comparison, we further fine-tune the pre-trained weights of the motion module provided by AnimateDiff and carefully adjust the hyperparameters. For ModelScopeT2V and LoRA fine-tuning, we train spatial and temporal parameters/LoRAs of the pre-trained video diffusion model for subject and motion respectively, and then merge them during inference. In addition, we also evaluate our generation quality for customizing subjects and motions independently. We evaluate our method against Textual Inversion [18] and Dreamix [50] for subject customization while comparing with Tune-A-Video [77] and ModelScopeT2V for motion customization.

**Evaluation metrics.** We evaluate our approach with the following four metrics. (1) *CLIP-T* calculates the average cosine similarity between CLIP [55] image embeddings of all generated frames and their text embedding. (2) *CLIP-I* computes the average cosine similarity between the CLIP image embeddings of all generated frames and target subject images. (3) *DINO-I* [57] measures the visual similarity between generated and target subjects using ViTS/16 DINO [8]. Compared to CLIP, the self-supervised model encourages distinguishing features of individual subjects. (4) *Temporal Consistency* [17], we compute CLIP image embeddings on all generated frames and report the average cosine similarity between all pairs of consecutive frames.

### 4.2. Results

In this section, we showcase results for both joint customization as well as individual customization of subjects and motions, further demonstrating the flexibility and effectiveness of our method.

**Arbitrary combinations of subjects and motions.** We compare our DreamVideo with several baselines to evalu-

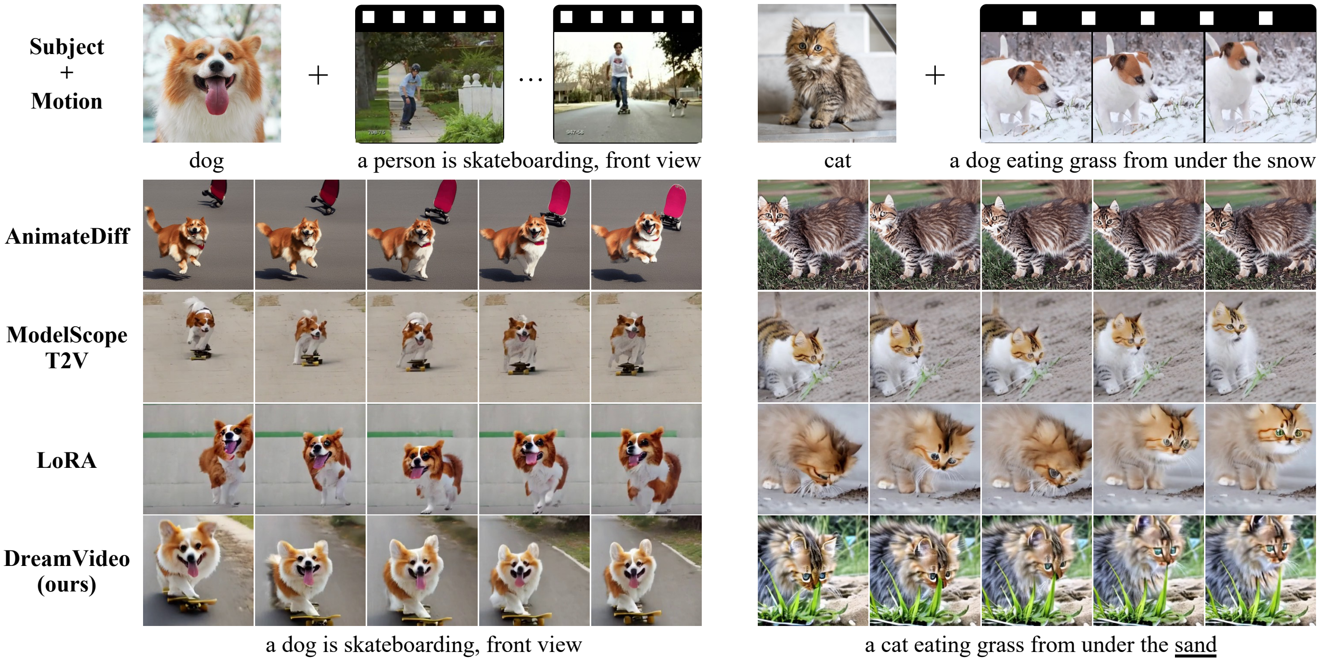

Subject
+
Motion
dog
a person is skateboarding, front view
cat
a dog eating grass from under the snow
AnimateDiff
ModelScope T2V
LoRA
DreamVideo (ours)
a dog is skateboarding, front view
a cat eating grass from under the sand

Figure 5. **Qualitative comparison of customized video generation with both subjects and motions**. DreamVideo accurately preserves both subject identity and motion pattern, while other methods suffer from fusion conflicts to some extent. Note that the results of AnimateDiff are generated by fine-tuning its provided pre-trained motion module and appending it to a DreamBooth [57] model.

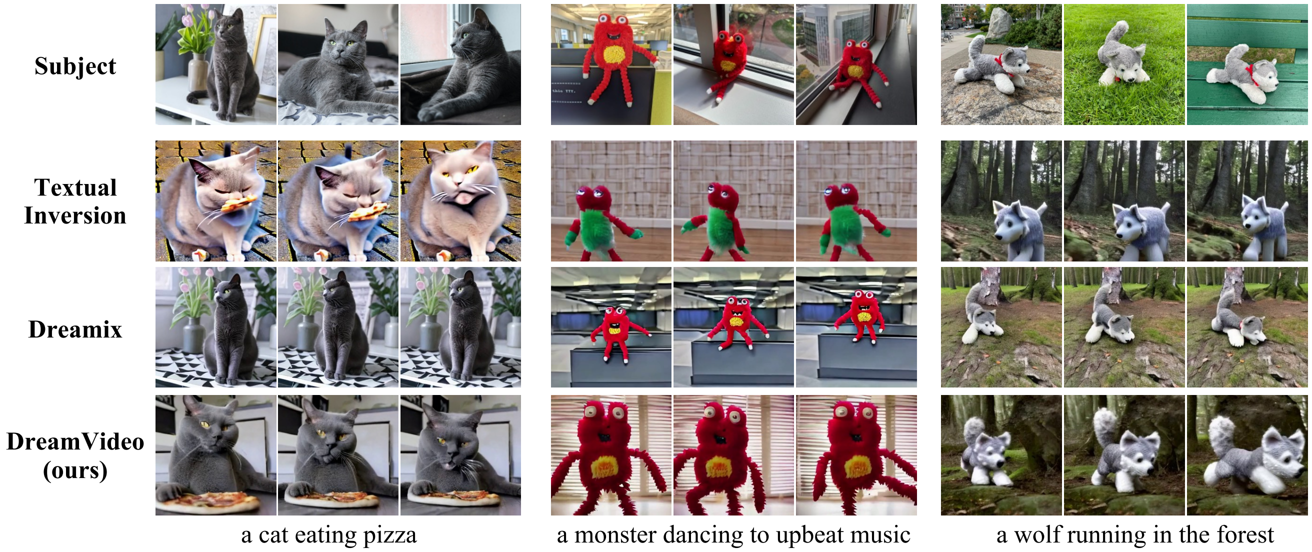
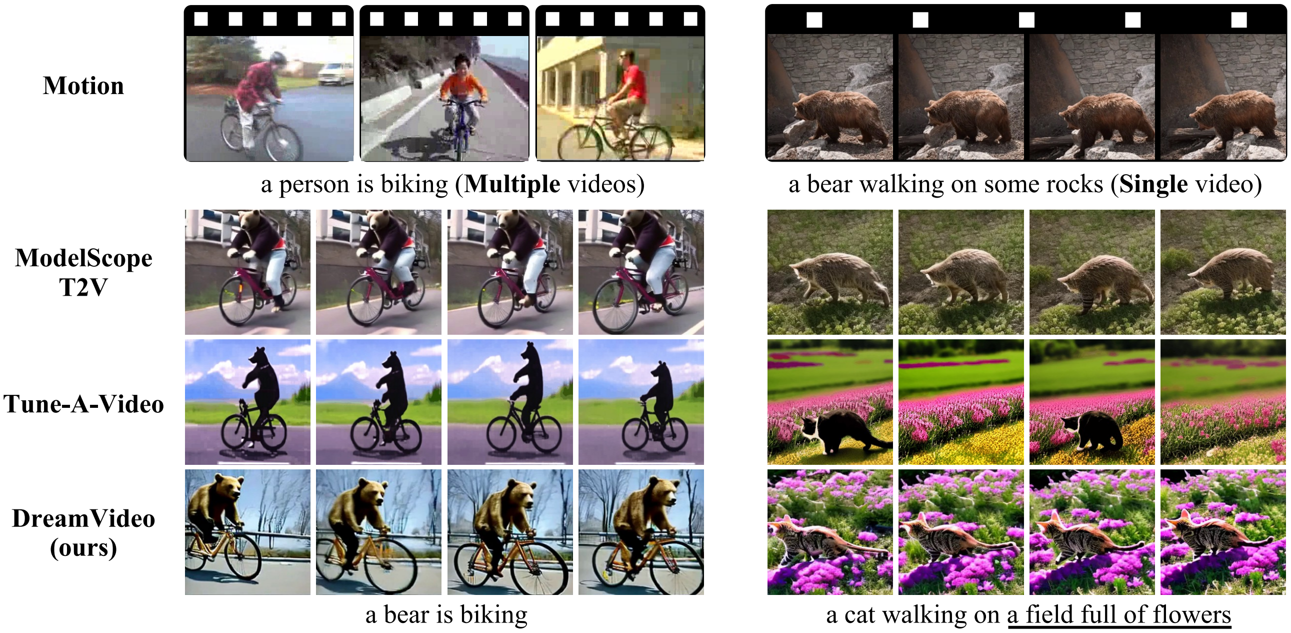


Figure 6. **Qualitative comparison of subject customization**. Our DreamVideo generates customized videos that preserve the precise subject appearance while conforming to text prompts with various contexts.

ate the customization performance, as depicted in Fig. 5. We observe that AnimateDiff preserves the subject appearances but fails to model the motion patterns accurately, resulting in generated videos lacking motion diversity. Furthermore, ModelScopeT2V and LoRA suffer from fusion conflicts during combination, where either subject identities are corrupted or motions are damaged. In contrast, our DreamVideo achieves effective and harmonious combinations that the generated videos can retain subject identities and motions under various contexts.

Tab. 1 shows quantitative comparison results of all methods. DreamVideo outperforms other methods across CLIP-



a person is biking (**Multiple** videos)

a bear walking on some rocks (**Single** video)

a bear is biking

a cat walking on a field full of flowers

Figure 7. **Qualitative comparison of motion customization** between DreamVideo and other methods. Our approach effectively models specific motion patterns while avoiding appearance coupling, and generates temporal coherent as well as diverse videos.

| Method | CLIP-T | CLIP-I | DINO-I | T. Cons. | Para. |
|---|---|---|---|---|---|
| AnimateDiff [23] | 0.298 | 0.657 | 0.432 | **0.982** | 1.21B |
| ModelScopeT2V [70] | 0.305 | 0.620 | 0.365 | 0.976 | 1.31B |
| LoRA | 0.286 | 0.644 | 0.409 | 0.964 | 6M |
| **DreamVideo** | **0.314** | **0.665** | **0.452** | 0.971 | 85M |

Table 1. **Quantitative comparison of customized video generation by combining different subjects and motions.** "T. Cons." denotes Temporal Consistency. "Para." means parameter number.

T, CLIP-I, and DINO-I, which is consistent with the visual results. Although AnimateDiff achieves highest Temporal Consistency, it tends to generate videos with small motions. In addition, our method remains comparable to Dreamix in Temporal Consistency but requires fewer parameters.

**Subject customization.** To verify the individual subject customization capability of our DreamVideo, we conduct qualitative comparisons with Textual Inversion [18] and Dreamix [50], as shown in Fig. 6. For a fair comparison, we employ the same baseline model, ModelScopeT2V, for all compared methods. We observe that Textual Inversion makes it hard to reconstruct the accurate subject appearances. While Dreamix captures the appearance details of subjects, the motions of generated videos are relatively small due to overfitting. Moreover, certain target objects in the text prompts, such as "pizza" in Fig. 6, are not generated by Dreamix. In contrast, our DreamVideo effectively mitigates overfitting and generates videos that conform to text descriptions while preserving precise subject appearances.

The quantitative comparison for subject customization is shown in Tab. 2. Regarding the CLIP-I and Temporal Consistency, our method demonstrates a comparable performance to Dreamix while surpassing Textual Inversion. Remarkably, our DreamVideo outperforms alternative methods in CLIP-T and DINO-I with relatively few parameters. These results demonstrate that our method can efficiently model the subjects with various contexts.

| Method | CLIP-T | CLIP-I | DINO-I | T. Cons. | Para. |
|---|---|---|---|---|---|
| Textual Inversion [18] | 0.278 | 0.668 | 0.362 | 0.961 | 1K |
| Dreamix [50] | 0.284 | **0.705** | 0.459 | **0.965** | 823M |
| **DreamVideo** | **0.295** | 0.701 | **0.475** | 0.964 | 11M |

Table 2. **Quantitative comparison of subject customization.**

**Motion customization.** Besides subject customization, we also evaluate the motion customization ability of our DreamVideo by comparing it with several competitors, as shown in Fig. 7. For a fair comparison, we only fine-tune the temporal parameters of ModelScopeT2V to learn a motion. The results show that ModelScopeT2V inevitably fuses the appearance information of training videos, while Tune-A-Video suffers from discontinuity between video frames. In contrast, our method can capture desired motion patterns while ignoring the appearance of training videos, generating temporally consistent and diverse videos.

As shown in Tab. 3, our DreamVideo achieves the highest CLIP-T and Temporal Consistency compared to baselines, verifying the superiority of our method.

**User study.** To further evaluate our approach, we conduct

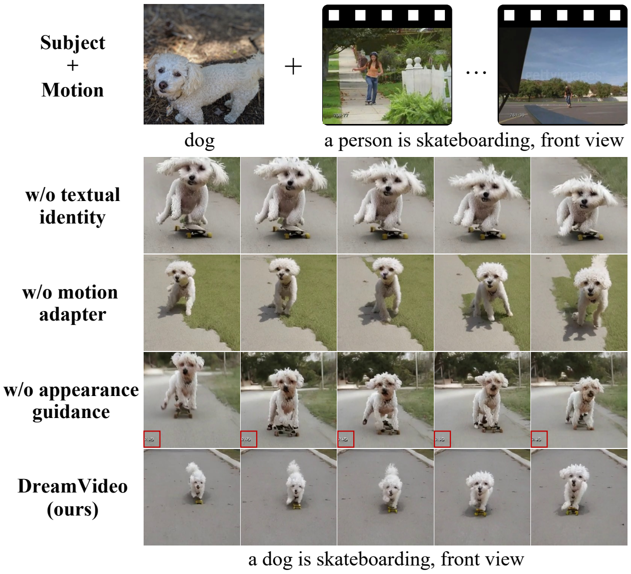


Figure 8. **Qualitative ablation studies** on each component.

| Method | CLIP-T | T. Cons. | Para. |
|---|---|---|---|
| ModelScopeT2V [70] | 0.293 | 0.971 | 522M |
| Tune-A-Video [77] | 0.290 | 0.967 | 70M |
| **DreamVideo** | **0.309** | **0.975** | 74M |

Table 3. **Quantitative comparison of motion customization.**

| Method | Text Alignment | Subject Fidelity | Motion Fidelity | T. Cons. |
|---|---|---|---|---|
| ours *vs.* AD [23] | 72.3 / 27.7 | 62.3 / 37.7 | 82.4 / 17.6 | 64.2 / 35.8 |
| ours *vs.* MS [70] | 62.4 / 37.6 | 66.2 / 33.8 | 56.6 / 43.4 | 51.5 / 48.5 |
| ours *vs.* LoRA | 82.8 / 17.2 | 53.5 / 46.5 | 83.7 / 16.3 | 67.4 / 32.6 |

Table 4. **Human evaluations** on customizing both subjects and motions between our method and alternatives. "AD" and "MS" are short for AnimateDiff and ModelScopeT2V, respectively.

user studies for subject customization, motion customization, and their combinations respectively. For combinations of specific subjects and motions, we ask 5 annotators to rate 50 groups of videos consisting of 5 motion patterns and 10 subjects. For each group, we provide 3∼5 subject images and 1∼3 motion videos; and compare our DreamVideo with three methods by generating videos with 6 text prompts. We evaluate all methods with a majority vote from four aspects: Text Alignment, Subject Fidelity, Motion Fidelity, and Temporal Consistency. Text Alignment evaluates whether the generated video conforms to the text description. Subject Fidelity and Motion Fidelity measure whether the generated subject or motion is close to the reference images or videos. Temporal Consistency measures the consistency between video frames. As shown in Tab. 4, our approach is most preferred by users regarding the above four aspects.

| Method | CLIP-T | CLIP-I | DINO-I | T. Cons. |
|---|---|---|---|---|
| w/o textual identity | 0.310 | 0.657 | 0.435 | 0.968 |
| w/o motion adapter | 0.295 | **0.701** | **0.475** | 0.964 |
| w/o appearance guidance | 0.305 | 0.650 | 0.421 | 0.970 |
| DreamVideo | **0.314** | 0.665 | 0.452 | **0.971** |

Table 5. **Quantitative ablation studies** on each component.

### 4.3. Ablation Studies

We conduct an ablation study on the effects of each component in the following.

**Effects of each component.** As shown in Fig. 8, we can observe that without learning the textual identity, the generated subject may lose some appearance details. When only learning subject identity without our devised motion adapter, the generated video fails to exhibit the desired motion pattern due to limitations in the inherent capabilities of the pre-trained model. In addition, without appearance guidance, the subject identity and background in generated videos may be corrupted due to the coupling of spatial and temporal information. These results demonstrate each component makes contributions to the final performance.

The quantitative results in Tab. 5 show that all metrics decrease slightly without textual identity or appearance guidance, illustrating their effectiveness. Furthermore, we observe that only customizing subjects leads to the improvement of CLIP-I and DINO-I, while adding the motion adapter can increase CLIP-T and Temporal Consistency. This suggests that the motion adapter helps to generate temporal coherent videos that conform to text descriptions.

## 5. Conclusion

In this paper, we present DreamVideo, a novel approach for customized video generation with any subject and motion. DreamVideo decouples video customization into subject learning and motion learning to enhance customization flexibility. We combine textual inversion and identity adapter tuning to model a subject and train a motion adapter with appearance guidance to learn a motion. With our collected dataset that contains 20 subjects and 30 motion patterns, we conduct extensive experiments, demonstrating the efficiency and flexibility of our method in both joint customization and individual customization of subjects and motions.

**Limitations.** Although our method can efficiently combine a single subject and motion, it fails to generate videos containing multiple subjects with multiple motions. One possible solution is to train a general customized video model.

**Acknowledgement** This work was supported in part by National Natural Science Foundation of China (No. 62101136), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), ZJ Lab, Shanghai Center for Brain Science and Brain-inspired Technology, and Alibaba Group.

## 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

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-A-Scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2, 4

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

[4] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional GAN with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 2

[5] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 2

[6] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2LIVE: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 2

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 5

[9] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2Video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 2

[10] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[11] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. DisenBooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 2

[12] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3

[13] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 2

[14] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-A-Video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 2

[15] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[16] Zhongjie Duan, Lizhou You, Chengyu Wang, Cen Chen, Ziheng Wu, Weining Qian, Jun Huang, Fei Chao, and Rongrong Ji. DiffSynth: Latent in-iteration deflickering for realistic video synthesis. *arXiv preprint arXiv:2308.03463*, 2023. 2

[17] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2, 5

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 4, 5, 7

[19] 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 *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 2

[20] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2

[21] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent diffusion fea-

tures for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

[22] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-Show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 2, 5

[23] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 5, 7, 8

[24] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[25] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022. 2

[26] 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. 2

[27] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 4

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[30] 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

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 3

[32] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 2

[33] 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

[34] 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. 3, 5

[35] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-Bloom: Zero-shot text-to-video generator with LLM director and LDM animator. *arXiv preprint arXiv:2309.14494*, 2023. 2

[36] Zhizhong Huang, Junping Zhang, and Hongming Shan. When age-invariant face recognition meets face age synthesis: A multi-task learning framework and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 2

[37] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2

[38] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2, 4

[39] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. CCVS: context-aware controllable video synthesis. *Advances in Neural Information Processing Systems*, 34:14042–14055, 2021. 2

[40] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. VideoGen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2

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

[42] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 4

[43] 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. 2

[44] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 2

[45] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 2, 5

[46] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 2

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 2

[48] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2

[49] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Deng Zhidong. DreamTalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 2

[50] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2, 3, 5, 7

[51] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. ST-Adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 3

[52] 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. 5

[53] 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. *arXiv preprint arXiv:2303.09535*, 2023. 2

[54] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 4, 5

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[57] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 5, 6

[58] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. HyperDreamBooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[59] Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MoStGAN-V: Video generation with temporal motion styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5661, 2023. 2

[60] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. InstantBooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[61] 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. *arXiv preprint arXiv:2209.14792*, 2022. 2

[62] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 2

[63] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual Diffusion: Continual customization of text-to-image diffusion with C-LoRA. *arXiv preprint arXiv:2304.06027*, 2023. 2

[64] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 5

[65] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2015. 5

[66] 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. 5

[67] 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

[68] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018.

[69] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[70] 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. 2, 3, 5, 7, 8

[71] 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

[72] 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. 2, 3

[73] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. *arXiv preprint arXiv:2312.15770*, 2023. 2

[74] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. VideoLCM: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023.

[75] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. LAVIE: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2

[76] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[77] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 5, 8

[78] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. LAMP: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 2

[79] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2

[80] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. SimDA: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 3

[81] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[82] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023. 2

[83] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. 3

[84] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[85] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. InstructVideo: Instructing video diffusion models with human feedback. 2023. 2

[86] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2

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

[88] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[89] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 2

[90] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. MotionDirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 3

[91] 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