VideoBooth: Diffusion-based Video Generation with Image Prompts

Yuming Jiang[1] Tianxing Wu[1] Shuai Yang[3]* Chenyang Si[1]
Dahua Lin[2] Yu Qiao[2] Chen Change Loy[1] Ziwei Liu[1]✉
[1]S-Lab, Nanyang Technological University [2]Shanghai AI Laboratory
[3]Wangxuan Institute of Computer Technology, Peking University
https://vchitect.github.io/VideoBooth-project/

Figure 1. **Videos synthesized by image prompts**. Our VideoBooth generates videos with the subjects specified in the image prompts.

Abstract

*Text-driven video generation witnesses rapid progress. However, merely using text prompts is not enough to depict the desired subject appearance that accurately aligns with users' intents, especially for customized content creation. In this paper, we study the task of video generation with image prompts, which provide more accurate and direct content control beyond the text prompts. Specifically, we propose a feed-forward framework **VideoBooth**, with two dedicated designs: **1)** We propose to embed image prompts in a coarse-to-fine manner. Coarse visual embeddings from image encoder provide high-level encodings of image prompts, while fine visual embeddings from the proposed attention injection module provide multi-scale and detailed encoding of image prompts. These two complementary embeddings can faithfully capture the desired appearance. **2)** In the attention injection module at fine level, multi-scale image prompts are fed into different cross-frame attention layers as additional keys and values. This extra spatial information refines the details in the first frame and then it is propagated to the remaining frames, which maintains temporal consistency. Extensive experiments demonstrate that VideoBooth achieves state-of-the-art performance in generating customized high-quality videos with subjects specified in image prompts. Notably, VideoBooth is a generalizable framework where a single model works for a wide range of image prompts with only feed-forward passes.*

*This work was done when Shuai Yang was in S-Lab, NTU.

1. Introduction

Text-to-image models [16, 17, 20, 26, 31, 37, 39, 43–45, 62, 67–69, 73, 75] have attracted substantial attention. With Stable Diffusion [69], we can now easily generate images using texts. Recently, the focus has been shifted to text-to-video models [6, 18, 30, 38, 46, 48, 60, 76, 80, 84, 85, 97] to generate videos by taking text descriptions as inputs. However, in some user cases, texts alone are not expressive enough to define the specific appearance of subjects [21, 71]. For example, as shown in Fig. 2, if we want to generate a video clip containing the dog as the third row, we need to use several attributive adjuncts to define the appearance of the dog in text prompts. Even with these extensive attributive adjuncts, models still cannot generate the desired appearance. Defining the appearance of the desired object with texts alone has the following flaws: 1) It is hard to enumerate all the desired attributes, and 2) The model cannot capture all attributes accurately with a long text. Compared to using texts, a more straightforward way to define the appearance is to provide reference images, termed image prompts. The image prompts are complementary to the text prompts and enrich the details that are hard to be depicted by text prompts.

There are several attempts to introduce image prompts into text-to-image models, which can be roughly divided into two groups. One is to fine-tune parts of parameters using few-shot reference images [15, 21, 28, 51, 71], which contain the same objects captured under different circumstances. However, the requirement for the number of reference images is demanding as sometimes it is not practical to obtain multiple images of the same object. The other category [11, 42, 52, 83, 87, 88], aiming to address this limitation, proposes to embed image prompts into text-to-image models and the inference is tuning-free. Both of these two types of attempts achieve plausible results in generating images containing objects specified in image prompts.

In this paper, we study a more challenging task, *i.e.*, text-to-video generation with image prompts. The task has two main challenges: 1) Similar to text-to-image generation, the attributes of image prompts should be accurately captured and then reflected in the generated videos; 2) Different from text-to-image generation, we aim for the dynamic movement of the object rather than a static one. Directly adapting these methods to video domain results in mismatched appearance or unnatural degraded movements. To address these challenges, we proposed **VideoBooth** with elaborately designed coarse-to-fine visual embedding components: 1) Coarse visual embeddings via image encoder: An image encoder is trained to inject the image prompts into text embeddings; 2) Fine visual embeddings via attention injection: The image prompts are mapped to multi-scale latent representations to control the generation process through cross-frame attentions of text-to-video models.

Specifically, inspired by early attempts [42, 83] in text-to-image models with image prompts, we extract the CLIP image features of the provided image prompts using the pre-trained CLIP model [66]. Then the extracted features are mapped into the text embedding space, which are inserted to replace parts of the original text embeddings. The well-trained encoder embeds the coarse appearance information of the given image prompts. However, coarse visual embedding is a universal embedding: 1) It only contains high-level semantic information, and 2) It is shared across all blocks with the same scale. As a result, some visual details are missing in the coarse visual embeddings.

To further refine the generated details as well as maintain temporal consistency, different from the highly compacted coarse visual embeddings, multi-scale image prompts are injected into cross-frame attention modules in different layers. The image prompts provide spatial information as well as details with different granularities. On the one hand, keeping spatial information of the image prompts can retain more details. On the other hand, different cross-frame attention modules need detailed information at different scales. Specifically, the latent representations of image prompts are appended as additional keys and values to refine the details in the first generated frame. To propagate the refined first frame to the following frames to maintain temporal consistency, we then use the updated values of the first frame as values for the remaining frames.

We set up a dedicated VideoBooth dataset to support the study of the new task. With each video, we provide an image prompt and a text prompt. Extensive experiments demonstrate the effectiveness of our proposed VideoBooth to generate videos with subjects specified in image prompts. As shown in Fig. 1, videos generated by VideoBooth better keep the visual attributes of image prompts. Besides, our proposed VideoBooth is tuning-free at inference time and videos can be generated with feed-forward passes only. The contributions are summarised as follows:

- To our knowledge, we are the first to explore the task of video generation using image prompts without finetuning

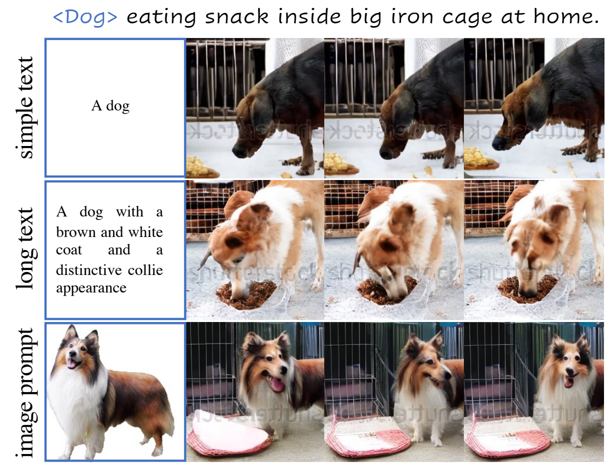


Figure 2. **The Use of Image Prompts.** We generate three video clips using different types of prompts: simple text prompt, long text prompt, and image prompt. We use the LLaVa model [55] to generate a text prompt describing the appearance of the image prompt. Using text prompts alone cannot fully capture the visual characteristics of the image prompt.

at inference time. We propose a dedicated dataset to support the task. Our proposed **VideoBooth** framework can generate consistent videos containing the subjects specified in image prompts.

- We introduce a new coarse-to-fine visual embedding strategy by image encoder and attention injection, which better captures the characteristics of the image prompts.
- We propose a novel attention injection method, using the multi-scale image prompts with spatial information to refine the generated details.

2. Related Work

Text-to-Video Models take the text descriptions as inputs and generate clips of videos. Early explorations [34, 80] on text-to-video models are based on the idea of VQVAE. Make-A-Video [76] proposes to add temporal attention to the architecture of DALLE2 model [68]. Recently, the emergence of diffusion models [32, 70] boosts research on text-to-video models [1, 23, 30, 33, 60, 82, 97]. Video LDM [6] proposes to train the text-to-video models on Stable Diffusion with temporal attention and 3D convolution introduced to handle the temporal generation. Gen-1 [18] introduces depth maps to handle the temporal consistency of text-to-video models. Some methods [27, 95] resort to training separate modules for synthesizing motions. All of the methods initialize their models with pre-trained text-to-image models. Another paradigm of using text-to-image models is to directly apply Stable Diffusion [69] to few-shot or zero-shot settings. Tune-A-Video [84] adapts the self-attention into cross-frame attention and then finetunes the stable diffusion model on a video clip. Models trained in this way have the capability to transfer motions from original videos. Text2Video-Zero [48] proposes to generate videos by using correlated noise maps to improve consistency. Apart from video generation, diffusion models have been applied to video-to-video generations [8, 9, 19, 24, 36, 41, 54, 58, 64, 65, 74, 81, 89, 90, 94].

Customized Content Creation aims at generating images and videos using reference images [15]. For customized text-to-image generation, optimization-based methods [28, 35, 51, 53] are proposed to optimize the weights of the diffusion model. For example, Textual Inversion [21] optimizes the word embeddings, while DreamBooth [71] proposes to finetune the weights of Stable Diffusion as well. Optimization-based methods require several reference images with the same subject to avoid the overfitting of the model, which is demanding in real-world applications. The cost of finetuning hampers the practical usage of these methods. To address these limitations, encoder-based methods [61, 88, 91, 98] are proposed to learn a mapping network to embed the reference images. ELITE [83] proposes to learn a global mapping network and local mapping network to encode the images into word embeddings. Jia *et al*. [42] propose to use an additional cross-attention to embed the image features. With the trained encoder, the personalized generation can be achieved in a feed-forward pass. Some recent works [2, 11, 22, 25] combine the encoder-based model and finetuning-based model to improve the performance. BLIP-Diffusion [52] proposes to pretrain a multimodal encoder in a large-scale dataset and then finetune the model on the specific subject for inference. Customized image generation is also applied to place the objects into the user-specified scenes [4, 13, 50, 77, 93]. Also, some efforts [12, 14, 40, 72, 79, 86, 87, 92] have been made to personalized face generation. He *et al*. [29] propose to improve the performance from the data perspective. Some works [3, 47, 59] focus on composing multiple subjects in one image. Apart from works on image generation, there are some early attempts at personalized video manipulation. Make-A-Protagonist [96] edits an existing video in a personalized way using Stable Diffusion 2.1 to embed the image prompts. The motion of the original video is learned from Tune-A-Video [84]. VideoDreamer [10] proposes to generate personalized videos by generating the first frames using a finetuning-based method and then generating the video clip using the Text2Video-Zero [48]. Different from existing works, our proposed VideoBooth does not need to finetune any weights at the inference time.

3. VideoBooth

Our proposed VideoBooth aims at generating videos from an image prompt I and a text prompt T. The image prompt specifies the appearance of the subject. An overview of our proposed VideoBooth is illustrated in Fig. 3. The image prompt is fed into VideoBooth in two levels. At the coarse level, it is fed into a pretrained CLIP Image encoder to extract visual features. An encoder, composed of several MLP layers, is trained to map visual features into the space of text embeddings. The obtained embedding f_I will be inserted into text embedding, which is extracted by feeding text prompt T into CLIP text encoder. To further refine the synthesized details, we propose to inject image prompt I into the cross-frame attention module in the pretrained video diffusion model. Specifically, we append latent representation x_I^t of image prompt I into the cross-frame attention. In this way, multi-scale visual details with spatial information are involved in the calculation of attention maps so that visual characteristics can be better preserved. Two ways of feeding image prompt corporate with each other in a coarse-to-fine manner. The encoder provides coarse visual embeddings of the image prompt, while the attention injection provides fine visual embeddings.

3.1. Preliminary: Pretrained Text-to-Video Model

Our proposed VideoBooth is developed based on the pretrained text-to-video model [30, 82, 84]. In this section, we

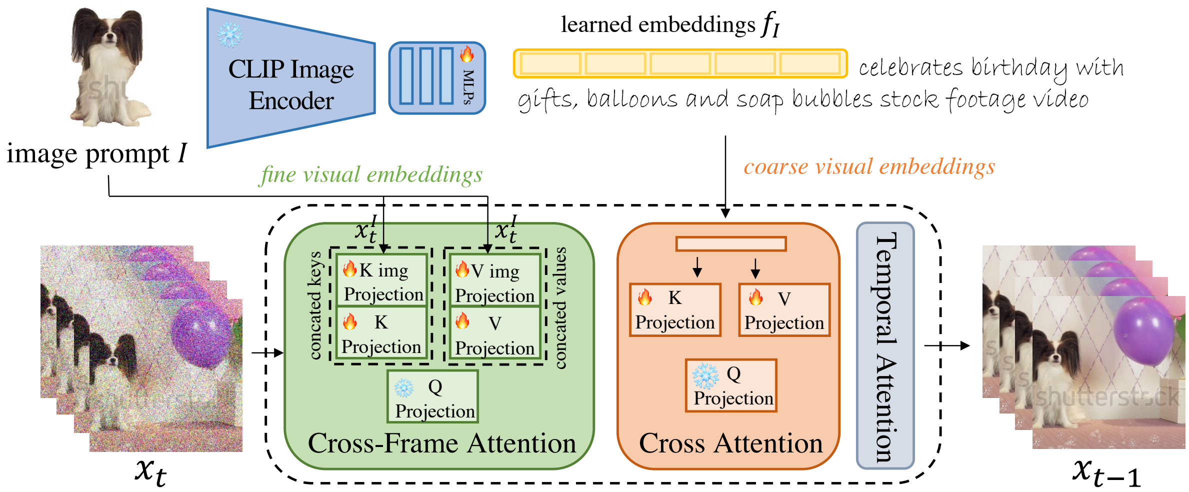


Figure 3. **Overview of VideoBooth.** VideoBooth generates videos by taking image prompts I and text prompts T as inputs. The image prompt is fed into the CLIP image encoder, followed by MLP layers. The obtained coarse visual embedding f_I is then inserted into the text embeddings. The composed embeddings serve as the input for cross attention. The embedding extracted by the encoder provides a coarse encoding of the visual appearance of the image prompt. To further refine the details in the generated videos, at the fine level, we append the latent representation of the image prompt to the cross-frame attention as additional keys and values. Different cross-frame attention layers receive latent representations with different scales. The multi-scale features with spatial details refine the synthesized details.

will briefly introduce the framework of text-to-video model.

Inflated 2D Conv. To handle video data and capture the temporal correlation, 2D conv in the Stable Diffusion model is inflated to 3D conv. In this way, the U-Net can encode 3D features containing the temporal dimension.

Cross-Frame Attention Module. Stable Diffusion has a self-attention module, where the features are enhanced by attending to themselves. To improve temporal consistency, the self-attention is modified into cross-frame attention. Specifically, the feature of each frame is enhanced by attending and referencing to the first frame and the previous frame. The cross-frame attention operates on both the spatial domain and temporal domain, thus the temporal consistency of the synthesized frames is improved.

Temporal Attention Module. Apart from cross-frame attention, a temporal attention module is introduced to further improve temporal consistency. Temporal attention operates on temporal domain and attends to all frames.

3.2. Coarse Visual Embeddings via Image Encoder

Given an image prompt I and text prompt T, the generated video is supposed to be consistent with visual elements and textual elements. Inspired by previous attempts at image-based customization methods [42, 83], we propose to encode visual information of image prompts by an image encoder. The image prompt and text prompt complement each other. The image prompt provides visual characteristics of the desired subject in the video, and the text prompt provides other orthogonal information. The extracted visual embeddings are combined with text embeddings as the final embeddings for the cross-attention module. Specifically, the CLIP image encoder is employed to extract the visual features f_V of image prompt I. Since the discrepancy exists between the CLIP image and text embeddings, f_V is then fed into MLP layers $F(\cdot)$ to map f_V to the spaces of text embeddings. The final embedding f_I for the image prompt is obtained as follows:

$$f_V = \text{CLIP}_I(I), f_I = F(f_V). \tag{1}$$

As for the text prompt T, we feed it into the CLIP text encoder to extract the text embedding f_T:

$$f_T = [f_T^0, f_T^1, ..., f_T^k, ...], \tag{2}$$

where f_T^k is the k-th word embedding in the text prompt.

To make the diffusion model generate videos conditioning on both text prompts and image prompts, we need to integrate these two embeddings, *i.e.*, f_I and f_T. The idea is to replace the word embedding of the target subject with f_I. Mathematically, f_I and f_t are fused to obtain the final text condition c_t as follows:

$$c_T = [f_T^0, f_T^1, ..., f_T^{k-1}, f_I, f_T^{k+n}, ...], \tag{3}$$

where k is the token index of the target subject in the text embedding, and n is the length of the text tokens for the target subject. For example, an image of a papillon dog is provided as an image prompt. To fuse the information from the image prompt and the text prompt "Papillon dog celebrates birthday with gifts", the word embeddings of the

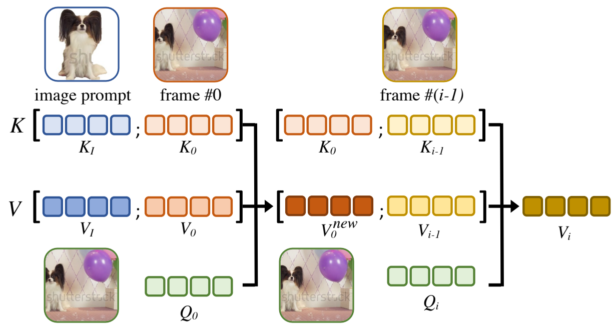


Figure 4. **Fine Visual Embedding Refinement.** We propose to inject the latent representation of image prompt (here we use the image for illustration purpose) directly into the cross-frame attention module. We use the keys and values from the image prompt to update the values of the first frame firstly. Then, the updated values of the first frame are used to update the remaining frames. Injecting the image prompt in the cross-frame attention helps to transfer the detailed visual characteristics of the image prompts to the synthesized frames. We perform the refinement in different cross-attention layers with different scales.

"papillon dog" will be replaced with f_I before they are fed into the cross-attention module of the diffusion models.

During the training of the coarse stage, we fix the parameters of the CLIP image encoder, and train the MLP layers. To make the diffusion model accommodate with the composed text embeddings c_T, we also finetune K and V projections (linear layers to map the input feature to the corresponding keys and values) in the cross-attention module.

3.3. Fine Visual Embeddings via Attention Injection

The well-trained image encoder embeds the coarse visual embeddings for image prompts and thus the synthesized videos contain the subjects specified in image prompts. However, the image encoder projects the image prompt into a flattened high-level representation, resulting in the loss of its detailed visual cues. Thus, some detailed visual characteristics in the image prompts may not be well preserved. To address this problem, a more effective way to preserve these details is to provide the model with the image prompts with spatial resolutions.

To further refine the synthesized details, we propose to inject image prompts into the cross-frame attention of the text-to-video models. By injecting image prompts into the cross-frame attention, the image prompts are involved in the updates of the synthesized frames so that the model can directly borrow some visual cues from image prompts.

Since text-to-video diffusion models operate in the latent space, we first feed the image prompt into the VAE of Stable Diffusion and get its latent representation x_I. Moreover, since the sampling of the videos starts from the noise map, the latent in the intermediate timesteps contains the noises. If we append the clean latent x_I of the image prompt to the cross-frame attention, the domain discrepancy exists. Therefore, we follow the diffusion forward process to add corresponding noises to x_I:

$$x_t^I = \sqrt{\overline{\alpha}_t}x_0^I + \sqrt{1-\overline{\alpha}_t}\epsilon, \tag{4}$$

where $\overline{\alpha}$ is a hyperparameter determined by the denoising schedule and $\epsilon \sim N(0, I)$.

The cross-frame attention is used to improve the temporal consistency of the generated frames. For each frame, the key and value are the concatenation of the features of the first frame and the previous frame. Here, we introduce the image prompts as the additional keys and values for the frames. As shown in Fig. 4, we propose to update the values of the first frame firstly using the keys and values of the image prompts and the frame itself. Mathematically, the operation can be expressed as follows:

$$\begin{aligned} V_0^{new} &= softmax(\frac{Q_0 K^T}{\sqrt{d}}) \cdot V, \\ K &= [K_I, K_0], V = [V_I, V_0], \end{aligned} \tag{5}$$

where K_I and V_I are the keys and values obtained from the image prompts. The query, key, and value of the first frame are denoted Q_0, K_0, and V_0, respectively. It should be noted that we use a separately trained K and V projection for latent representations x_t^I of image prompts because the image prompts have clean backgrounds, which are different from other frames. The parameters of the newly added K and V projections are initialized by original K and V projections.

Then the updated first frame is used to refine the remaining frames. When updating the remaining frames, the keys used for calculating the attention maps are the original keys, while the values are the updated ones. The update is expressed as follows:

$$\begin{aligned} V_i^{new} &= softmax(\frac{Q_i K^T}{\sqrt{d}}) \cdot V, \\ K &= [K_0, K_{i-1}], V = [V_0^{new}, V_{i-1}]. \end{aligned} \tag{6}$$

To sum up, in the attention injection, we update the values of the first frame using the image prompts first, and then use the updated first frame to update the other frames. In this way, the visual cues from the image prompts can be consistently propagated to all the frames.

It should be noted that the diffusion model has multiple cross-frame attention layers with different scales. To inject multi-scale visual cues for better detail refinement, in different cross-frame attention layers, we feed latent representations of the image prompts with corresponding resolutions, which are obtained from different stages of the U-Net.

3.4. Coarse-to-Fine Training Strategy

The visual details of the image prompts are embedded into the final synthesized results in two stages: coarse visual embeddings using an image encoder and fine visual embedding by attention injection. We propose to train these two modules in a coarse-to-fine manner. In other words, we train the coarse image encoder and tune the parameters in the cross-attention first. After the model has the capability of generating videos containing the subjects specified in image prompts, we then train the attention injection module to embed image prompts into cross-frame attention layers. As we will show in the ablation study (Sec. 5.5), if these two modules are trained together, the fine attention injection module leaks the strong visual cues and the coarse encoder learns meaningless representations. As a result, in sampling phase, the image encoder for the coarse visual embedding cannot provide the coarse information and then the fine attention module cannot refine the details. Therefore, it is necessary to train VideoBooth in a coarse-to-fine manner.

4. VideoBooth Dataset

We establish the VideoBooth dataset to support the task of video generation using image prompts. We start from the WebVid dataset [5], a well-known open-source dataset for text-to-video generation. In the WebVid dataset, there is a text prompt with each video. In this paper, we study the task of generating a video clip from one text prompt and one image prompt. Hence, in addition to the original text prompt, we need to provide an image prompt for each video. We propose to segment the subjects from the first frame of the video using the Grounded-SAM (Grounded Segment Anything) [49, 57], and the segmented subjects are image prompts. The Grounded-SAM receives word prompts as inputs and generates segmentation masks for the target subjects specified in word prompts. To obtain the word prompt for the input to Grounded-SAM, we use the spaCy library to parse the noun chunks from the original text prompts, which are used as the word prompts. After the segmentation, we perform data filtering to ensure the data quality. We filter out small objects and large objects (those are almost the same size as the original video) according to the ratio of the object to the whole video. Also, since we focus on generating video clips containing moving objects, we further filter the videos containing moving objects. The keywords we used for filtering are dog, cat, bear, car, panda, tiger, horse, elephant, and lion. In the current version, we have processed 2.5M subset of the WebVid dataset. After data filtering, we have 48,724 video data pairs for training. We will process the full set of the WebVid dataset and include the filtered data in our VideoBooth dataset.

To evaluate the performance, we also set up a test benchmark. The test benchmark consists of 650 test pairs. For each pair, an image prompt and a text prompt are provided. The test pairs are selected from the rest of the WebVid-10M dataset, which does not overlap with the training set.

Table 1. **Quantitative Comparisons.** VideoBooth achieves the best image alignment and comparable text alignment performance.

Method	CLIP-Text ↑	CLIP-Image ↑	DINO ↑
Textual Inversion [21]	29.9749	69.7995	45.3143
DreamBooth [71]	**30.6877**	71.2078	52.9661
ELITE [83]	30.0881	73.7518	58.9522
VideoBooth (Ours)	30.0967	**74.7971**	**65.0979**

5. Experiments

5.1. Comparison Methods

Textual Inversion [21] is a method for customized text-to-image generation. The appearance of the target subjects is embedded into the text embeddings. Concretely, a text token S^* is optimized to represent the subject. We adapt it to the task of text-to-video generation by replacing the image model with the video model.

DreamBooth [71] is also proposed for customized text-to-image generation. It injects the target subject into the text tokens as well as model weights. During the training, both the model weights and the special token $S*$ are optimized.

ELITE [83] is an encoder-based method for personalized generation. An encoder is trained to embed the images into the text embeddings. Local mapping and global mapping are employed to transform the CLIP embedding of image prompts into the features, which are injected into the cross-attention module. We adapt and retrain the method using the same pretrained video model we use.

5.2. Evaluation Metrics

We use three metrics to evaluate the performance [83]. To measure the alignment of the generated videos and given text prompts, we use the **CLIP-Text** metric. The metric is calculated using the cosine similarity of the CLIP text embeddings of text prompts and CLIP image embeddings of the generated frames. For each video, the value is obtained by averaging values of the all frames. As for the evaluation of the similarity between the given image prompts and the generated videos, we adopt two metrics: **CLIP-Image** and **DINO** [7, 63]. The CLIP-Image metric is calculated by the cosine similarity between the CLIP image embedding of image prompts and generated frames. Since the CLIP model is trained to align image embeddings and text embeddings, we follow the practice in previous methods [13, 71] and use the DINO similarity as another indicator. DINO is trained to differentiate the differences between objects of the same classes. We use the ViT-S/16 model to extract the features of the image prompts and generated frames. The final score is obtained by averaging over all frames.

5.3. Quantitative Comparisons

We report quantitative results in Table 1. As shown in Table 1, our proposed VideoBooth achieves state-of-the-art

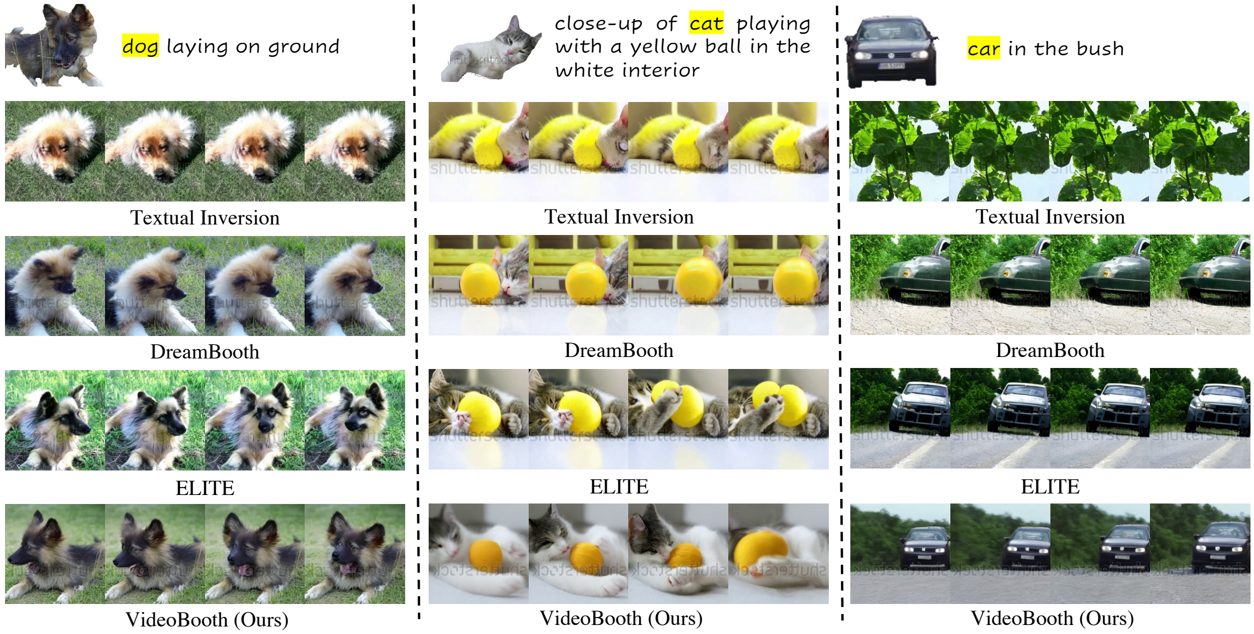


Figure 5. **Qualitative Comparison.** VideoBooth effectively preserves the fidelity of image prompts and achieves better visual quality.

image alignment performance compared to baseline methods. As for the alignment with the text prompts, our proposed VideoBooth has comparable performance with the baseline models. It should be noted that the CLIP-Text score of DreamBooth is significantly higher than the other methods. The reason lies in that the optimized token $S*$ in DreamBooth is inserted into the text embeddings, rather than replacing the original word embeddings like other methods. This would result in the generated videos of DreamBooth being highly related to the text prompts but having no correlation to the image prompts in some cases. We also conducted a user study, in which 25 users participated. Each user is presented with twelve groups of videos, and each group contains four videos generated by four methods. For each group, users are asked to make three choices: 1) which one has the best image alignment? 2) which one has the best text alignment? 3) which one has the best overall quality. Figure 6 summarizes the results. Our results are preferred by most users in all three dimensions.

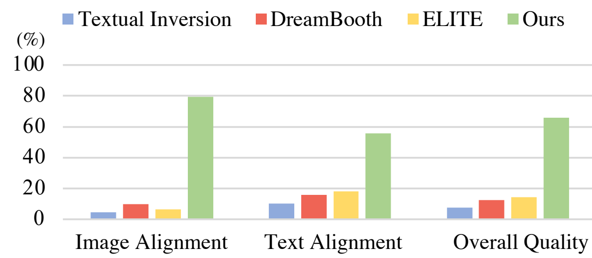


Figure 6. **User Study.** Our proposed VideoBooth achieves the highest user preference ratios on all three dimensions.

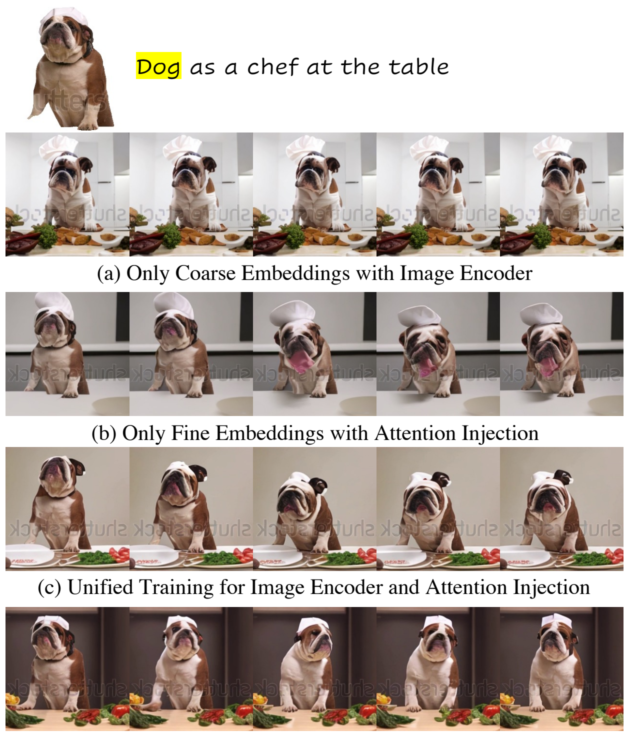
5.4. Qualitative Comparisons

We show three visual comparisons on the generated video frames of our proposed VideoBooth and baseline methods in Fig. 5. In the first example, the model is supposed to generate videos containing the dog specified in the image laying on the ground. Textual Inversion cannot correctly embed the appearance of the dog and results in generating another totally different dog. DreamBooth and ELITE can embed the coarse appearance of the dog but the generated details vary from the image prompt. Our proposed VideoBooth successfully embeds the details of the image prompts into the synthesized videos. In the second example, the condition is to generate a cat playing with the yellow ball. All the methods can generate a yellow ball and a cat but only our proposed VideoBooth can accurately generate the cat having the appearance from the image prompt. As for the last example, Textual Inversion fails to generate a car. DreamBooth generates a distorted car in the bush. ELITE model can generate a car in the bush, but the color differs from the image prompt. By contrast, our model can generate videos having the same car in the bush.

5.5. Ablation Study

To evaluate the effectiveness of the proposed components, we perform three ablation studies. Due to the computational resources, we train these models on the subset of our training set. The quantitative metrics are shown in Table 2.

Only Coarse Embeddings. This ablation model injects the image prompts with only coarse embeddings via Image Encoder. In the example shown in Fig. 7(a), the ablation model only encodes the coarse appearance of the image prompts. The pattern in the legs of the synthesized dog is different



(a) Only Coarse Embeddings with Image Encoder

(b) Only Fine Embeddings with Attention Injection

(c) Unified Training for Image Encoder and Attention Injection

(d) Full Model

Figure 7. **Ablation Study.** (a) With only coarse embeddings from image encoder, generated patterns in the body of dog are different from image prompt. (b) With only fine embeddings from attention injection, there lacks coarse encodings of the dog for attention injection module to refine and thus the generated dog is distorted at later synthesized frames. (c) The unified training degrades the capability of image encoder and thus the dog is distorted. (d) The full model better keeps the all visual details of image prompt.

from that in the image prompt. By contrast, the results of our full model show that our proposed model can transfer all the details in the image prompts to the synthesized videos.

Only Fine Embeddings. In this ablation model, we only have fine embeddings of the image prompt in cross-frame attention layers. The main purpose of using fine embedding is to refine the coarse encoding of image prompts from the Image Encoder. Without coarse embeddings, the model with fine embeddings only cannot refine the details. As shown in Fig. 7(b), the first frame contains the exact appearance as the image prompt, but the temporal consistency cannot be guaranteed. The generated dog is distorted in the following frames. The reason is that the model trained in this way overfits the image prompt. In the first frame, the model can copy the information from image prompts. In the following frames, without the coarse embeddings, the generation of the appearance only relies on the propagation of the appearance from the first frame.

The Necessity of Coarse-to-Fine Training. In VideoBooth, we propose the coarse-to-fine training strategy, *i.e.*, train the coarse embeddings first and then train the attention

Table 2. **Ablation Study.** The full model has the best scores.

Variants	**CLIP-Image ↑**	**DINO ↑**
(a) Coarse Embeddings only	75.4366	64.9568
(b) Fine Embeddings only	75.5553	66.0378
(c) Unified Training	75.8254	67.4201
(d) Full Model	**76.1631**	**69.7374**

injection module. In this ablation model, we train these two modules within one stage. The unified training makes the model rely heavily on the strong guidance provided in attention injection. In this way, the trained image encoder has limited capability. Thus, the model trained in this way has a similar behavior as the model with only fine embeddings. The model also overfits the image prompts from the attention injection. As shown in the first example in Fig. 7(c), the appearance in the first frame is correct, but the generated dog in the following frames is distorted. Due to the image encoder having limited capability to provide the correct coarse encoding of image prompts, the attention injection cannot refine the details. By contrast, the full model can generate consistent frames with all details well preserved.

6. Discussion

In this paper, we propose a novel framework VideoBooth to generate videos using image prompts and text prompts. The image prompts specify the appearance of the subjects. We inject the image prompts into the model in two modules: Coarse Embeddings via Image Encoder and Fine Embeddings via Attention Injection. The Image Encoder provides the coarse embeddings of image prompts for the refinement of the Attention Injection module. These two modules cooperate with each other and they are trained in a coarse-to-fine manner. Our proposed VideoBooth generates consistent videos containing the desired subjects.

Limitations and Future Work. The model is trained with the image prompts that do not have very divergent views from target videos. Thus, if an image prompt with a back view is provided, we cannot generate videos from the front view. In future work, we can augment image prompts using image-to-3D models [56, 78] so that image prompts used for training will have different views from target videos.

Potential Negative Societal Impacts. The model may be applied to generate fake videos, which can be potentially avoided by using deepfake detection methods.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This study is also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221-0012), NTU NAP, and the Google PhD Fellowship. This study is partially supported by the National Key R&D Program of China (2022ZD0160201), and Shanghai Artificial Intelligence Laboratory.

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. 3

[2] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06925*, 2023. 3

[3] 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. 3

[4] Jinbin Bai, Zhen Dong, Aosong Feng, Xiao Zhang, Tian Ye, Kaicheng Zhou, and Mike Zheng Shou. Integrating view conditions for image synthesis. *arXiv preprint arXiv:2310.16002*, 2023. 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 *IEEE International Conference on Computer Vision*, 2021. 6

[6] 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. 1, 3

[7] 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. 6

[8] Duygu Ceylan, Chun-Hao Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. *arXiv:2303.12688*, 2023. 3

[9] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023. 3

[10] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 3

[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, 3

[12] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023. 3

[13] 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. 3, 6

[14] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023. 3

[15] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*, 2023. 2, 3

[16] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1

[17] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 1

[18] 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. 1, 3

[19] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. *arXiv preprint arXiv:2309.16496*, 2023. 3

[20] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 1

[21] 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, 2022. 2, 3, 6

[22] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 3

[23] 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 *ICCV*, 2023. 3

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

[25] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*, 2023. 3

[26] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1

[27] 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. 3

[28] 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, 3

[29] Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Helge Rhodin, and Ratheesh Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. *arXiv preprint arXiv:2311.04315*, 2023. 3

[30] 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. 1, 3

[31] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

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

[33] 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. 3

[34] 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. 3

[35] 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. In *International Conference on Learning Representations*, 2022. 3

[36] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 3

[37] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6080–6090, 2023. 1

[38] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023. 1

[39] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 1

[40] Junha Hyung, Jaeyo Shin, and Jaegul Choo. Magicapture: High-resolution multi-concept portrait customization. *arXiv preprint arXiv:2309.06895*, 2023. 3

[41] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023. 3

[42] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023. 2, 3, 4

[43] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 1

[44] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[45] Yuming Jiang, Ziqi Huang, Tianxing Wu, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained 2d and 3d facial editing via dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[46] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

[47] Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Teare. An image is worth multiple words: Learning object level concepts using multi-concept prompt learning. *arXiv preprint arXiv:2310.12274*, 2023. 3

[48] 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. 1, 3

[49] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 6

[50] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17089–17099, 2023. 3

[51] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023. 2, 3

[52] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 2, 3

[53] Yuheng Li, Haotian Liu, Yangming Wen, and Yong Jae Lee. Generate anything anywhere in any scene. *arXiv preprint arXiv:2306.17154*, 2023. 3

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

[55] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[56] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3:

Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 8

[57] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6

[58] 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

[59] 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. 3

[60] 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. 1, 3

[61] 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. 3

[62] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1

[63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[64] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023. 3

[65] 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. 3

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[67] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[68] 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. 3

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 3

[70] 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. 3

[71] 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. 2022. 2, 3, 6

[72] 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. 3

[73] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

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

[75] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023. 1

[76] 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. 1, 3

[77] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 3

[78] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 8

[79] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. *arXiv preprint arXiv:2306.06638*, 2023. 3

[80] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 1, 3

[81] Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[82] 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. 3

[83] 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, 3, 4, 6

[84] 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. *arXiv preprint arXiv:2212.11565*, 2022. 1, 3

[85] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023. 2

[86] Zijie Wu, Chaohui Yu, Zhen Zhu, Fan Wang, and Xiang Bai. Singleinsert: Inserting new concepts from a single image into text-to-image models for flexible editing. *arXiv preprint arXiv:2310.08094*, 2023. 3

[87] 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, 3

[88] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking ”text” out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 2, 3

[89] Hanshu Yan, Jun Hao Liew, Long Mai, Shanchuan Lin, and Jiashi Feng. Magicprop: Diffusion-based video editing via motion-aware appearance propagation. *arXiv preprint arXiv:2309.00908*, 2023. 3

[90] 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

[91] 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

[92] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023. 3

[93] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*, 2023. 3

[94] 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. 3

[95] 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

[96] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023. 3

[97] 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, 3

[98] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023. 3