This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Fine-grained Controllable Video Generation via Object Appearance and Context

Hsin-Ping Huang[1,2] Yu-Chuan Su[1] Deqing Sun[1] Lu Jiang[1]
Xuhui Jia[1] Yukun Zhu[1] Ming-Hsuan Yang[1,2]

[1]Google DeepMind [2]University of California, Merced

Abstract

While text-to-video generation shows state-of-the-art results, fine-grained output control remains challenging for users relying solely on natural language prompts. In this work, we present FACTOR for fine-grained controllable video generation. FACTOR provides an intuitive interface where users can manipulate the trajectory and appearance of individual objects in conjunction with a text prompt. We propose a unified framework to integrate these control signals into an existing text-to-video model. Our approach involves a multimodal condition module with a joint encoder, control-attention layers, and an appearance augmentation mechanism. This design enables FACTOR to generate videos that closely align with detailed user specifications. Extensive experiments on standard benchmarks and user-provided inputs demonstrate a notable improvement in controllability by FACTOR over competitive baselines.

1. Introduction

Recent text-to-video models [6, 18, 25, 27, 32, 49, 50, 52, 59, 60, 80] allow users to translate their creative ideas into video content easily. However, achieving precise control over the composition of these videos remains a challenge. Specifying detailed object movements and appearances through text alone is often difficult and requires iterative revisions. Even when users provide additional descriptions like "from right to left" or "yellow and black trim," models can still struggle to generate the desired output, as shown in Fig. 1 top. It is highly desirable to have a user-friendly system that enables fine-grained control over the appearance and motion of individual objects.

Recent advances such as ControlNet [75] offer potential solutions, integrating structural controls (e.g., optical flow, depth) into text-to-video generation [10, 32, 61, 76]. Similar techniques have been developed in video editing [20, 38, 44, 54], where natural language guides the manipulation of a video's content and style. While these methods offer promising results, they rely on dense control inputs for every frame – typically extracted from a reference video (Fig. 1 middle). This limits the generation of novel videos with structures different from the reference and makes the process impractical for manual user input.

On the appearance control front, methods like DreamBooth and its extensions [6, 21, 32, 47, 61, 79] allow for subject customization in video generation. However, these methods are limited to adjusting the appearance of individual subjects and cannot manipulate their appearances and locations simultaneously. VideoComposer [56] takes a step forward by integrating text with spatial and temporal controls. However, it requires dense motion information from a reference video and applies these controls globally to the entire video. A method for generating videos composed of multiple entities, with precise control over appearance and motion, remains an exciting and open problem.

In this work, we introduce FACTOR, a framework for fine-grained controllable video generation. We demonstrate that for precise control, users should be able to easily manipulate individual entities when generating videos with multiple objects. FACTOR prioritizes user-friendliness by accepting sparse and intuitive inputs: a text prompt, user-drawn bounding boxes, and user-provided reference images. To this end, we build FACTOR upon an off-the-shelf text-to-video model [50]. We use a joint encoder to integrate the multimodal input control signals and insert control-attention layers into the model's transformer blocks. Training involves only these newly inserted layers, with the pre-trained model's weights frozen. This design ensures high-quality video generation while introducing object-level control mechanisms. We further enhance the diversity of reference appearance images through color and geometric transformations to improve appearance control. Experiments on the MSR-VTT benchmark [65] demonstrate that FACTOR significantly improves generation quality, trajectory, and appearance control compared to the alternative approaches. A user study further validates the effectiveness of the proposed method. The main contributions of this work are:

- We target the new form of *fine-grained controllable video generation* that aims to synthesize videos via multimodal context (text, appearance, and trajectory) of individual objects from easy-to-give user inputs.

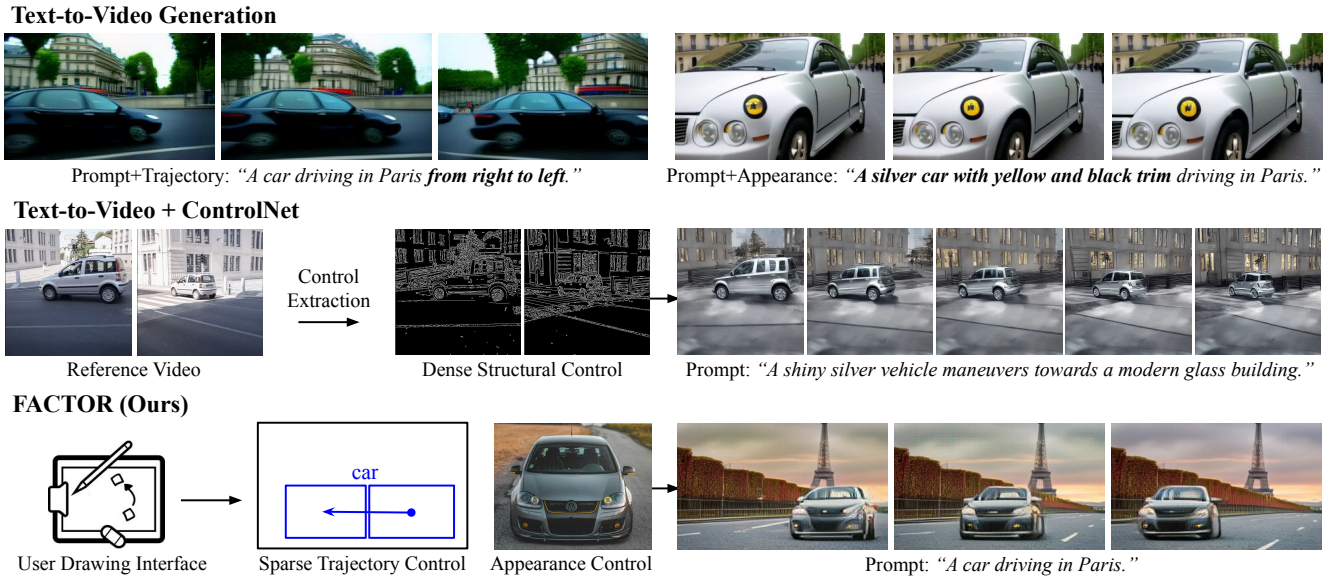


Figure 1. **Text-to-video generation** [50] has limited controllable ability through user-provided prompts. **Text-to-video+ControlNet** [76] requires dense control signals extracted from a reference video. To enable controllability through user-friendly inputs, **FACTOR** controls precise subject movements through hand-drawn trajectories and their visual appearance using reference examples.

- We propose a unified framework to achieve multi-modal control within a single training process. This is achieved by injecting control signals with control-attention and appearance augmentation.
- We validate that our method offers fine controllability of multiple objects compared to existing works and shows the additional benefit of creating interactions, which is challenging for existing text-to-video models.

2. Related Work

Text-to-video generation. Text-to-video models have made impressive progress. Token-based methods [27, 59, 60] utilize an auto-regressive model to predict videos in the latent space. Diffusion-based models [25, 26, 49] extend the 2D diffusion model [46] to generate videos [5, 6, 8, 9, 18, 19, 21, 22, 32, 33, 39, 52, 52, 53, 55, 57, 64, 74, 80] by incorporating temporal layers. Although these models produce promising results, they have limited control over the generated videos since text prompts cannot accurately convey precise control signals. In this work, we develop a fine-grained video generation framework that controls the location and appearance of objects. We adopt a generative transformer model [50] as our base model, and our approach can be adapted to diffusion-based models.

Controllable text-to-image generation. Various models have been proposed to enhance the controllability of text-to-image models, particularly in terms of structure and appearance. ControlNet [36, 42, 66, 71, 75, 78], structure-guided generation [1, 23, 67, 70, 73], and training-free approaches [3, 14, 34, 62, 72] incorporate various spatial control signals such as edges, depth, segmentation, and human pose into pre-trained diffusion models. However, the dense structure inputs these models require can be challenging for users to provide when generating videos. Fine-tuning on reference images [16, 35, 47] and encoder-based methods [17, 58] are employed to control the appearance of subjects. However, these methods do not control the locations of subjects. In contrast, we propose a fine-tuning-free method to jointly control the appearance and location of subjects for video generation. ControlNet is extended to control global appearance and style [66, 71], while FACTOR focuses on individual subject appearance control.

Controllable text-to-video generation. Structural control for video generation [10, 12, 13, 15, 32, 37, 40, 56, 61, 63, 76] focuses on producing temporally consistent videos using a temporal attention layer integrated with ControlNet. Video editing approaches [7, 20, 28, 30, 38, 43–45, 48, 51, 54, 64, 69] aim to alter the appearance and style of videos based on text prompts. They typically require dense structural inputs extracted from reference videos. In contrast, our work focuses on controllable T2V generation using sparse and user-friendly control signals. Video customization through fine-tuning [24, 32, 61, 79] still relies on dense structural inputs and often lacks control over subject location [6, 21]. VideoComposer [56] employs a reference image for global appearance control and a dense motion sequence for spatial control. In contrast, FACTOR focuses on controlling the generation of individual subjects using sparse inputs.

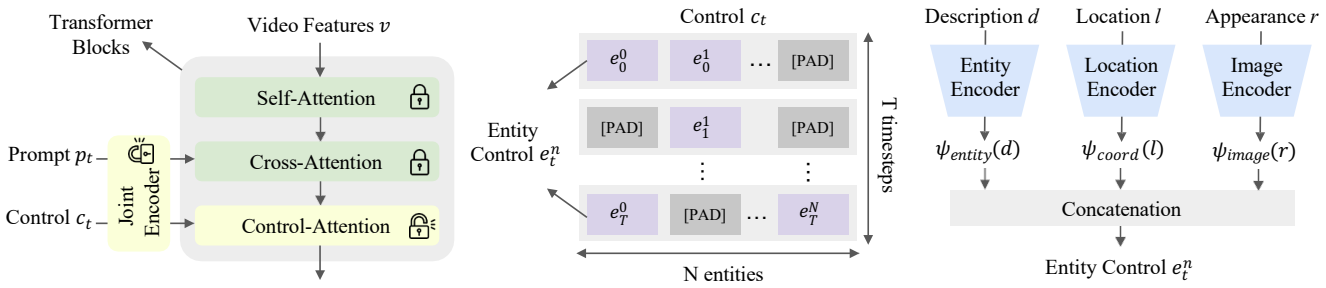


Figure 2. **Overview.** a) **Multimodal condition**: a joint encoder is learned to encode the prompt and control to capture their interaction. Control-attention layers are inserted into the transformer blocks of the text-to-video model to incorporate multimodal control signals. Only the inserted layers are optimized to generate videos satisfying the fine-grained control. b) **Entity control encoding**: given T time steps, the embedding of control c_t is formed by the control for N entities, e_t^n, where padding tokens replace the embedding of the non-existing entity. The control for entity n at time t is formed by embeddings of the description, location, and reference appearance.

3. Method

Our main goal is to develop a user-friendly system allowing entity-level motion and appearance control for video generation. To this end, we enable users to provide multimodal inputs: a text prompt, user-drawn bounding box trajectories, and user-provided reference images. By breaking down the video into manageable entities, users can create the desired video content by adjusting the properties of individual elements. Our intuitive control interface facilitates users to gradually design their videos by adding entities with reference images in one frame and adjusting their positions across multiple frames. We first briefly review the base T2V model and provide an overview of our approach. Then, we introduce each component in detail.

3.1. Preliminaries: Text-to-video Generation

Our base T2V model [50] consists of an encoder-decoder model that encodes the video into discrete tokens and a bidirectional transformer model that predicts the video tokens conditioned on the embedding of text prompts. During training, the tokens are replaced with a special token `[MASK]`, and the transformer model is optimized to predict the tokens at `[MASK]` locations based on the text embedding. We minimize the negative log-likelihood of predicting the masked tokens $v_t, t \in M$, where M denotes the subset of video tokens that are masked, v denotes the video token sequence, and v_M denotes the masked version of v.

$$\mathcal{L} = - \mathop{\mathbb{E}}_{v \in \mathcal{D}} \sum_{t \in M} \log p(v_t | v_M, p). \tag{1}$$

The transformer model contains a series of attention layers that condition the video token prediction on the text embeddings. At inference time, all tokens are replaced with `[MASK]`, and the model iteratively predicts the tokens. We use a token-based generative transformer model for its fast inference and inherent flexibility in modeling various control signals directly within the same architecture.

3.2. Overview

We present a novel approach for fine-grained video generation that allows for precise control of individual objects using user-friendly inputs. The problem is defined as follows: given a text prompt p and fine-grained control c as inputs, our model aims to generate a video that satisfies both input conditions. Specifically, users provide multimodal control c by 1) describing the desired entities in the video, 2) drawing their trajectories, and 3) providing a reference appearance image for each entity. This pipeline helps users create videos intuitively. The input control signal is given at T timesteps. Assuming there are N entities in the video, we define our control as $c = \{c_t\}_{t=1}^{T}$. At a single timestep t, the embeddings c_t are formed by a sequence of entity controls, i.e., the embeddings of the N entities $c_t = \{e_t^n\}_{n=1}^{N}$. The embeddings e_t^n encode the desired conditions to generate the entity indexed by n at time t, as shown in Fig. 2.

3.3. Multimodal Condition

We introduce an effective multimodal condition method for generating videos that satisfy fine control of different modalities within a single training pipeline. Our method includes a joint encoder and a control-attention module.

Joint encoder. Existing text-to-image models (*e.g.*, [36, 75]) use a separate encoder for each input condition and directly input the independent embeddings into the model, which is less effective (see Tab. 1 GLIGEN [36], VideoComposer [56]) as the encoder needs to be trained for new conditions and the interaction between the controls is ignored. In contrast, we use a joint encoder that simultaneously encodes the text prompt p and the fine-grained control c of different modalities within the same encoder. This design facilitates the learning of interactions between multimodal inputs, captured in the contextualized embeddings.

Control-attention. The next task is to incorporate the sequence of control embeddings into the base T2V model. A natural design choice is to include the control embed-

dings in the original cross-attention module. However, this poses difficulties in generating videos aligned with the new control because the cross-attention layers are optimized for the text prompt, weakening their flexibility for new inputs (see Tab. 1 ELITE [58]). Instead, we insert a new control-attention layer in each transformer block to accommodate additional control. During training, we freeze the weights of the pre-trained model and train the joint encoder and the new control-attention layer. This allows the T2V model to generate videos aligned with both text and the new multi-modal condition. By fixing the pre-trained weights, we preserve the capability to generate high-quality videos while updating only 23% of the parameters.

Our control-attention has less computational overhead (19%) compared to gated-attention [36], which struggles to converge with the increasing length of tokens in videos. While we use a generative transformer model, our control-attention layers can be extended to other T2V models, like diffusion models, which have similar attention blocks [6, 46]. This module can also handle other control signals, such as camera poses. We achieve this by transforming the control into tokens and inputting them into our joint encoder and control-attention layers. Our unified training process conditions the model on multimodal control signals in a single pipeline, avoiding the need for multiple condition-specific methods.

3.4. Entity Control Encoding

Here, we discuss our entity-level fine-grained control. The entity embeddings e_t^n are constructed by encoding the context of each entity. First, the description of the entity d is given by text, e.g., a cat, and encoded into embeddings $\psi_{\text{entity}}(d)$. Second, the location of the entity l is given by the top-left and bottom-right bounding box coordinates of the entity and encoded as $\psi_{\text{coord}}(l)$. Finally, the reference appearance r of the entity is given by a single example image and encoded by a CLIP image encoder as $\psi_{\text{image}}(r)$. The embeddings e_t^n are the concatenation of the description, location, and appearance embeddings of the entity:

$$e_t^n = \text{Concat}(\psi_{\text{entity}}(d_t^n), \psi_{\text{coord}}(l_t^n), \psi_{\text{image}}(r_t^n)). \quad (2)$$

We replace the embeddings of e_t^n with padding embeddings when the entity of index n is missing at timestep t.

User-friendly inputs. To make our method user-friendly, we assume descriptions and appearances are fixed throughout the video, and only the location changes over time although all the conditions can be thoroughly given at T timesteps. Our method takes sparse inputs where the location of the entity is provided by simply drawing a bounding box in the first frame and dragging it to move to the location in the last frame. We obtain the locations in middle frames by linear interpolation. Our sparse input is more user-friendly compared to [56], which is only applicable to dense inputs. In addition, we randomly drop the description, the location, and the appearance embeddings of the entities 20% of the time during training. In this case, the model is trained to generate results when one or more control signals are absent. At inference time, users can provide partial control signals as inputs (see Fig. 11).

Data collection. In practice, very few video datasets include annotations for object trajectories and visual examples. To train our model, we employ an off-the-shelf object detector [29] and tracking algorithm [4] to extract N entities within a video clip and their locations across T timesteps. For each entity, we collect a reference visual example r by cropping the region defined by the detected bounding boxes.

3.5. Appearance Augmentation

Unlike image synthesis [35, 47], it is difficult to collect multiple images of the same subject as training data for videos. Here, we discuss how we create diverse reference images of the subject. During training, we sample the reference visual example of the subject from a longer video clip outside our training clip within the same video to obtain reference images with more diverse appearances. However, these samples still contain limited backgrounds and poses. To create more diverse references, we apply strong augmentation to the reference images. Specifically, we use color augmentation to create diverse backgrounds and geometric augmentation to introduce different subject poses. These operations force the model to learn the subject's appearance and exclude irrelevant information such as backgrounds and poses, thus preventing the model from overfitting the generated results to the given reference image.

To further enhance the motion of live subjects at inference time, we simulate different poses of the subject by sampling a sequence of transformed variants of the reference image using translation and cropping. Specifically, we randomly crop the reference image to an initial appearance r_0^n. Then, we define a translation vector δ to augment the initial appearance. We create a sequence of reference images with enhanced poses, $r_t^n = T_{\delta \cdot t}(r_0^n)$, where T denotes the translation, as conditions at different timesteps. This strategy greatly improves the motion dynamics of generated live subjects, such as animals (See Tab. 3). In addition, the test-time augmentation can be prompt-dependent, such as using a translation vector spanning downward to customize the target motion of getting up (See Fig. 10).

4. Experiments

Dataset. Our model is trained on a dataset consisting of 10M videos from the WebVid dataset [2] and a randomly sampled subset of 500M images from the WebLI dataset [11]. Each training batch comprises 20% images and 80% videos.

Evaluation metrics. 1) `FVD` assesses video quality. 2)

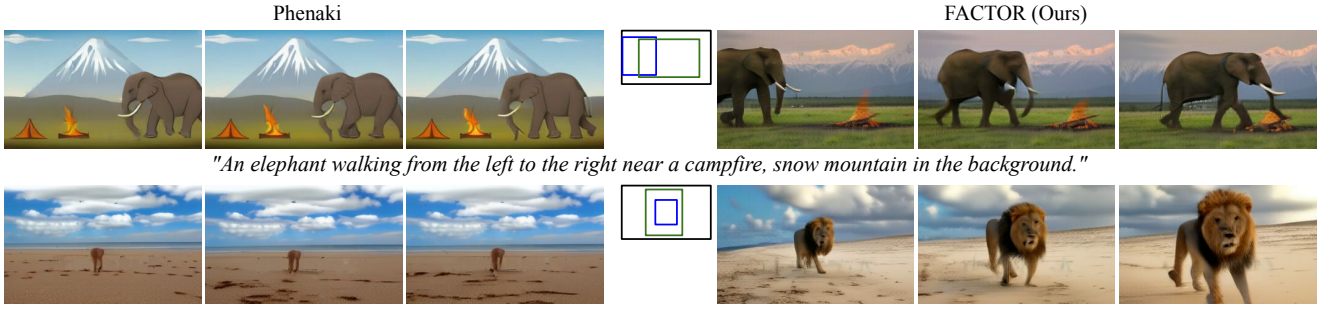


Figure 3. **Trajectory control.** We assess FACTOR's ability to control trajectories compared to the base model Phenaki [50]. The text prompt is augmented with trajectory descriptions as inputs to Phenaki. While Phenaki fails to accurately generate object movements, FACTOR successfully controls object movements. The blue and green boxes denote the positions in the initial and final frames.

Figure 4. **Enhanced subject interaction through trajectory control.** By utilizing trajectories of the two main entities as inputs, FACTOR improves the generation of subject interactions, a challenging task for T2V models.

`CLIP-T` measures the alignment between prompts and generated videos using CLIP embeddings. 3) `AP` evaluates alignment to trajectories by comparing detected bounding boxes in generated videos with ground truth using an object detector [29]. 4) `CLIP-V` quantifies alignment to reference images by comparing image similarity in generated videos with ground truth using CLIP embeddings [47].

Implementation details. We implement the base T2V model following the architecture of Phenaki [50]. The base model and FACTOR are trained for 1M and 500K steps with batch sizes of 256 and 128, respectively. Videos are generated at a length of 11 and a resolution of 192×320. The dimension of c_t is 220×512. The model is trained with a maximum of four entities. Further implementation details and model architectures are in the supplementary materials.

4.1. Trajectory Control

First, we evaluate FACTOR's ability to control the trajectories of generated entities. Since FACTOR is constructed based on Phenaki [50], we first validate that FACTOR enhances fine controllability compared to its base model without compromising quality. We design several prompts using descriptions of bounding box trajectories as inputs to Phenaki [50]. In Fig. 3, Phenaki fails to generate object movements aligned with the designed prompts, whereas FACTOR accurately generates the correct movement direction for the entity by conditioning the generation on hand-drawn trajectories. In Fig. 4, we use trajectories of two main entities as inputs. FACTOR generates videos aligned with these trajectories and additionally benefits by generating interactions between entities, such as "high-fiving", even though our method is not specifically trained on annotated interactions.

In Fig. 5, we demonstrate FACTOR's advantages by comparing it with state-of-the-art T2V and controllable T2V models. By conditioning the generation on input trajectories, FACTOR synthesizes videos with enhanced semantic meaning and larger motion, such as an astronaut stretching their hand to a duck. In contrast, the T2V model VideoLDM [6], which relies solely on text prompts, generates videos with less dynamic motion. We compare FACTOR with controllable T2V models [56, 68, 77] designed to control subject movements. VideoComposer requires a reference image and dense motion sequence, but with FACTOR's sparse trajectory inputs, it often generates static or fade-out results due to its reliance on dense inputs. Direct-a-Video, which controls object locations by amplifying attention maps, struggles with overlapping boxes. FACTOR better aligns objects to bounding boxes by training on a large-scale video dataset. MotionDirector, which requires reference *videos* for motion control, only approximately controls object locations and cannot generate structures different from the reference (e.g., an elephant lifting an object).

4.2. Appearance Control

We demonstrate FACTOR's capability of appearance control. For this evaluation, we use images of subjects

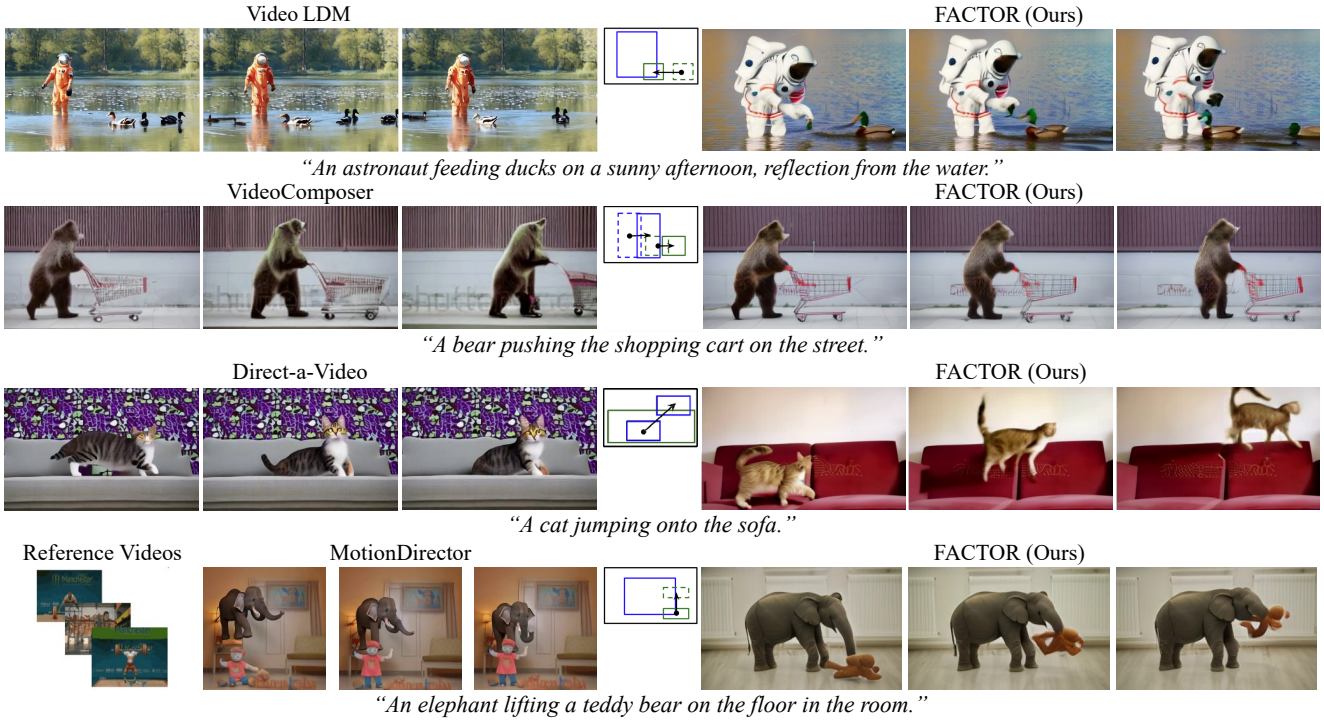


Figure 5. **Comparison to state-of-the-art.** By conditioning video generation on trajectories, FACTOR synthesizes videos with larger motions compared to T2V models. When using FACTOR's trajectory inputs, VideoComposer generates fade-out effects, Direct-a-Video struggles with overlapping boxes, and MotionDirector fails to generate videos with structures differing from the reference video.

Figure 6. **Comparison of subject customization.** FACTOR generates subjects with higher fidelity and larger movements compared to DreamBooth and VideoBooth, which lack trajectory control. MotionDirector struggles to generate accurate environments (e.g., New York).

from [35, 47]. Each subject's appearance is conditioned on a single reference image for FACTOR. In this section, we compare FACTOR to state-of-the-art subject customization methods. We further evaluate FACTOR's control injection by comparing it to the control modules developed for T2I models. We then show FACTOR's ability to achieve controllable image animation, a special type of appearance control.

First, we compare FACTOR with subject customization images and video synthesis methods including DreamBooth [6], VideoBooth [31], and MotionDirector [77]. Notably, these models cannot control entity trajectories or handle multiple entities. Fig. 6 shows FACTOR generates high-fidelity customized subjects and exhibits larger movements by controlling trajectories, outperforming DreamBooth and VideoBooth. MotionDirector [77] uses separate temporal and spatial LoRA for motion and appearance controls, resulting in inadequate motion and incorrect environments (e.g., New York).

Next, we evaluate FACTOR's control module by comparing it to the control modules of ELITE and GLIGEN [36, 58] implemented on our video backbone. Since no existing models achieve the same control on videos as FACTOR, we compare methods originally designed for T2I

Figure 7. **Comparison of control modules.** FACTOR generates videos that align accurately with input trajectories and appearances. Other control methods, ELITE and GLIGEN, often generate inaccurate subject appearances and struggle to control their locations.

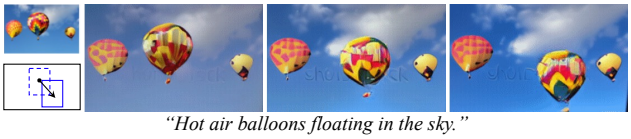


Figure 8. **Controllable image animation.** FACTOR achieves image animation conditioned on the initial frame and the trajectory.

tasks. In Fig. 7, we show that FACTOR generates subjects that closely align with input trajectories and appearances. In contrast, ELITE and GLIGEN often generate inaccuracies in subject appearance and struggle to effectively control their locations. GLIGEN's use of separate encoders for different controls introduces higher computational overhead and presents convergence challenges for video models, limiting its ability to inject precise control signals compared to FACTOR's unified encoder approach.

Though beyond the scope of this paper, we show FACTOR's capability for appearance control in image animation, aiming to generate videos conditioned on the first frame. In Fig. 8, FACTOR animates the hot air balloon in the initial frame based on the input trajectory. This capability leverages generative transformers adept at various infilling tasks. FACTOR's training involves predicting video tokens within random masks, enabling it to predict subsequent frames from the initial frame and control inputs, thus achieving controllable image animation.

4.3. Quantitative Results

We evaluate our models on the MSR-VTT test set [65], comprising 2,990 examples. We generate one video per example by randomly selecting one prompt from 20 prompts [6]. We use an unseen set of 6,513 videos as the real videos for FVD evaluation and compare two variants: 1) `FACTOR-T`: our model with trajectory control. 2) `FACTOR`: our model with trajectory and appearance control.

In Tab. 1, we first compare FACTOR and FACTOR-T with the base T2V model Phenaki, all using the same video backbone but accepting different levels of control inputs. To test our control module's capability, we use inputs automatically extracted from ground truth videos for FACTOR and FACTOR-T. This allows us to quantitatively evaluate the alignment between generated and ground truth videos, noting the extracted inputs are noisier than user-provided ones. FACTOR improves FVD scores with control inputs that enforce closer alignment to real video distributions in trajectories and appearances. However, it achieves slightly lower CLIP-T scores than Phenaki, likely due to misalignment between extracted controls and text prompts in test data. Longer training enhances control alignment but could reduce prompt alignment, as noted in concurrent works [41]. We also assess AP scores. While Phenaki is not designed for generating videos with specific object trajectories, its AP score reflects chance alignment when prompts sufficiently match input trajectories. FACTOR's higher AP score indicates successful alignment with provided trajectory controls. Additionally, FACTOR outperforms Phenaki in CLIP-V scores, showing superior alignment with appearance controls. We also present results from other state-of-the-art T2V models trained on different data and backbones for reference.

Table 1. **Quantitative results.** FACTOR achieves higher AP and CLIP-V scores, demonstrating its capability to generate videos aligned with trajectory and appearance control inputs.

Methods	FVD ↓	CLIP-T ↑	AP ↑	CLIP-V ↑
T2V models				
MagicVideo	1290	—	—	—
VideoLDM	—	0.2929	—	—
Make-A-Video	—	**0.3049**	—	—
ModelScope	550	0.2930	—	—
Controllable T2V models				
VideoComposer	342	0.2906	0.126	0.742
Direct-a-Video	360	0.2849	0.144	0.658
GLIGEN	217	0.2817	0.173	0.703
ELITE	130	0.2712	0.137	0.733
Same backbones				
Phenaki*	411	0.2870	0.099	0.663
FACTOR-T (Ours)	339	0.2787	0.290	0.683
FACTOR (Ours)	**116**	0.2721	**0.356**	**0.763**

*We implement and train the Phenaki model from scratch using our datasets.

Next, we compare FACTOR's control injection module with controllable T2V models. Compared to VideoComposer [56] and Direct-a-Video [68], FACTOR achieves bet-

Table 2. **User study.** FACTOR consistently achieves performance gains using inputs provided solely by users. Average scores on a 0-2 scale are reported based on 16 prompts and 5 subjects.

	Quality↑	Text↑	Trajectory↑	Appearance↑
Phenaki *vs.* FACTOR-T	1.13 / **1.63**	0.72 / **1.97**	0.16 / **1.96**	-
Phenaki *vs.* FACTOR	1.25 / **1.63**	1.70 / **1.75**	0.43 / **1.88**	0.38 / **1.75**

Table 3. **Ablation study.** FACTOR consistently outperforms ablated models across various metrics.

Methods	FVD↓	CLIP-T↑	AP↑	CLIP-V↑
FACTOR	**116**	0.2721	**0.356**	**0.763**
w/o joint encoder	133	0.2719	0.307	0.726
w/o control-attention	127	0.2720	0.322	0.750
w/o augmentation	124	**0.2723**	0.341	0.758

"Milk poured into a sks vase."

Figure 9. **Ablation study.** FACTOR achieves better alignment with control inputs compared with ablated models.

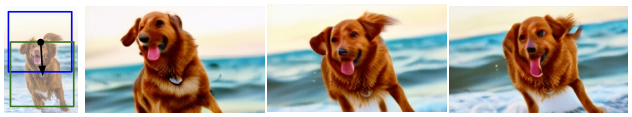
ter FVD, AP, and CLIP-V scores. These results highlight FACTOR's enhanced controllability with user-friendly trajectory inputs. Compared to ELITE and GLIGEN [36, 58], FACTOR also outperforms in terms of FVD, AP, and CLIP-V scores, highlighting the effectiveness of FACTOR's control-attention module.

Finally, we conduct a user study using input controls provided by users. Raters assess two videos generated by different methods using the same inputs on a 0-2 scale for the following criteria: quality, text alignment, trajectory alignment, and appearance alignment. Average scores from 16 videos and 5 subjects are reported in Tab. 2. FACTOR consistently achieves higher ratings across all criteria.

4.4. Ablation Study

We conduct an ablation study to validate the effectiveness of each proposed module. In Tab. 3 and Fig. 9, we compare the following models: 1) `w/o joint encoder`: uses separate encoders for prompt and fine control. 2) `w/o control-attention`: concatenates prompt and fine control inputs to the original cross-attention layer. 3) `w/o augmentation`: trained without appearance augmentation. FACTOR consistently achieves higher metrics compared to these alternative models, confirming the importance of the proposed components. Here, our appearance augmentation technique proves effective across different prompts, with augmentations randomly sampled.

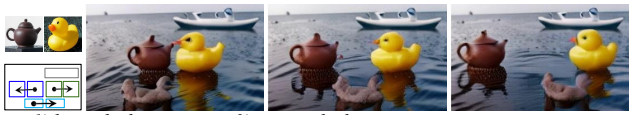
In Fig. 10, we further demonstrate that appearance augmentation can be dynamically adjusted at test time to better meet user preferences. By applying a translation vector, such as moving downwards for a jumping motion, we use variations of the reference image as appearance conditions



"A sks dog jumping on the beach."

Figure 10. **Appearance augmentation.** We demonstrate fine-tuned augmentation at test time as a form of control. For instance, adjusting the translation of the reference image downwards to achieve a desired jumping motion.

1) painting: no control, 2) cat: trajectory, 3) sofa: trajectory, appearance



1) boat, duck: trajectory, 2) teapot, duck toy: trajectory, appearance

Figure 11. **Partial and multiple control.** FACTOR demonstrates capability with partial inputs and supports up to four objects.

at different time steps, which enhances the realism of live subjects' motions. In Fig. 11, we show that FACTOR handles partial inputs where one or more controls are absent for the entities. Also, FACTOR supports up to four objects.

5. Conclusions

We introduce a novel approach for fine-grained, controllable video generation. Our framework allows users precise control over video creation by specifying entity names, drawing trajectories, and providing visual appearance examples. Key components include a joint encoder for learning interaction among controls, control-attention modules for precise control injection, and appearance augmentation for diverse subject poses. This approach enables nuanced control over the appearance and trajectory of multiple objects in generated videos.

Limitations. First, FACTOR's performance relies on the capabilities of the base T2V model. Integrating the proposed control module into more advanced models could enhance overall quality. Second, FACTOR implicitly manages camera poses (Fig. 6). Explicit control over camera poses could be achieved by integrating them into FACTOR's control sequences. Third, FACTOR faces challenges when text prompts do not align well with fine-grained control inputs. Lastly, generating subject poses significantly different from a single reference image is challenging for FACTOR. Training FACTOR on datasets with multiple references and upgrading the base model could improve its performance.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *CVPR*, 2023. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 4

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 4

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[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 *CVPR*, 2023. 1, 2, 4, 5, 6, 7

[7] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. *arXiv preprint arXiv:2308.09592*, 2023. 2

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024. 2

[9] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv preprint arXiv:2312.04557*, 2023. 2

[10] 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. 1, 2

[11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 4

[12] Yutao Chen, Xingning Dong, Tian Gan, Chunluan Zhou, Ming Yang, and Qingpei Guo. Eve: Efficient zero-shot text-based video editing with depth map guidance and temporal consistency constraints. *arXiv preprint arXiv:2308.10648*, 2023. 2

[13] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. *arXiv preprint arXiv:2308.10079*, 2023. 2

[14] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2023. 2

[15] 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. 2

[16] 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

[17] 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. *arXiv preprint arXiv:2302.12228*, 2023. 2

[18] 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. 1, 2

[19] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. *arXiv preprint arXiv:2305.10474*, 2024. 2

[20] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 1, 2

[21] 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. 1, 2

[22] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 2

[23] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. In *SIGGRAPH*, 2023. 2

[24] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 2

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2

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

[27] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 1, 2

[28] 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. 2

[29] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 4, 5

[30] Nisha Huang, Yuxin Zhang, and Weiming Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. *arXiv preprint arXiv:2305.05464*, 2023. 2

[31] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, 2024. 6

[32] 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. In *ICCV*, 2023. 1, 2

[33] 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. *ICCV*, 2023. 2

[34] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023. 2

[35] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2, 4, 6

[36] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 2, 3, 4, 6, 8

[37] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 2

[38] 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. 1, 2

[39] 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. *arXiv preprint arXiv:2303.08320*, 2023. 2

[40] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 2

[41] Joanna Materzyńska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2023. 7

[42] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[43] 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. 2

[44] 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. 1, 2

[45] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv preprint arXiv:2305.12328*, 2023. 2

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

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 2, 4, 5, 6

[48] 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. 2

[49] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1, 2

[50] 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. In *ICLR*, 2023. 1, 2, 3, 5

[51] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 2

[52] 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. 1, 2

[53] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 2

[54] 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. 1, 2

[55] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap at-

tention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2024. 2

[56] 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. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 7

[57] 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

[58] 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. In *ICCV*, 2023. 2, 4, 6, 8

[59] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 1, 2

[60] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 1, 2

[61] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 1, 2

[62] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023. 2

[63] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 2

[64] 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. 2

[65] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1, 7

[66] 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

[67] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *CVPR*, 2023. 2

[68] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM TOG*, 2024. 5, 7

[69] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, 2023. 2

[70] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. 2

[71] 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. 2

[72] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *ICCV*, 2023. 2

[73] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, and M. Patel, Vishal. Scenecomposer: Any-level semantic image synthesis. In *CVPR*, 2023. 2

[74] 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

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 2, 3

[76] 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. 1, 2

[77] 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. 5, 6

[78] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 2

[79] 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. 1, 2

[80] 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. 1, 2