This ICCV 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.

MagicMotion: Controllable Video Generation with Dense-to-Sparse Trajectory Guidance

Quanhao Li[1,2*] Zhen Xing[1,2*] Rui Wang[1,2] Hui Zhang[1,2] Qi Dai[3] Zuxuan Wu[1,2†]

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3] Microsoft Research Asia

https://quanhaol.github.io/magicmotion-site/

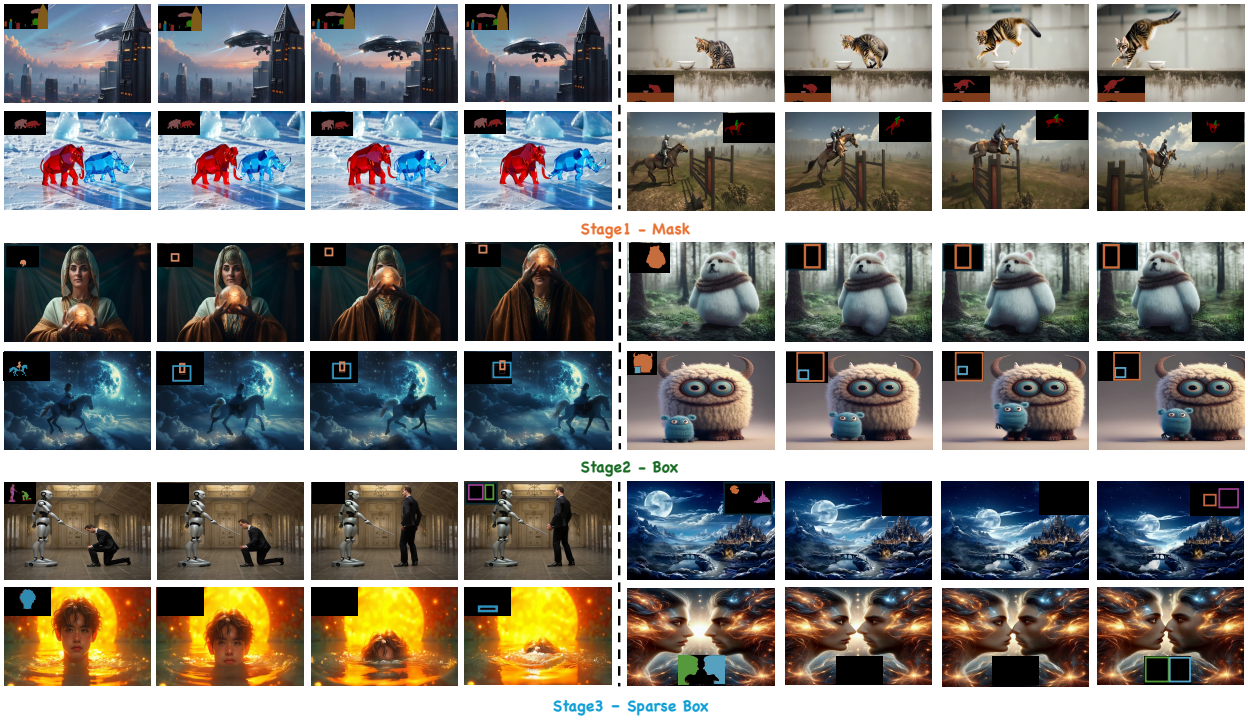


Figure 1. **Example videos generated by MagicMotion.** MagicMotion consists of three stages, each supporting a different level of control from dense to sparse: mask, box, and sparse box. Given an input image and any form of trajectory, MagicMotion can generate high-quality videos, animating objects in the image to move along the user-specified path.

Abstract

Recent advances in video generation have led to remarkable improvements in visual quality and temporal coherence. Upon this, trajectory-controllable video generation has emerged to enable precise object motion control through explicitly defined spatial paths. However, existing methods struggle with complex object movements and multi-object motion control, resulting in imprecise trajectory adherence, poor object consistency, and compromised visual quality. Furthermore, these methods only support trajectory control in a single format, limiting their applicability in diverse scenarios. Additionally, there is no publicly available dataset or benchmark specifically tailored for trajectory-controllable video generation, hindering robust training and systematic evaluation. To address these challenges, we introduce ***MagicMotion****, a novel image-to-video generation framework that enables trajectory control through three levels of conditions from dense to sparse: masks, bounding boxes, and sparse boxes. Given an input image and trajectories, MagicMotion seamlessly animates objects along defined trajectories while maintaining object consistency and visual quality. Furthermore,*

* equal contributions. † Corresponding author.

*we present **MagicData**, a large-scale trajectory-controlled video dataset, along with an automated pipeline for annotation and filtering. We also introduce **MagicBench**, a comprehensive benchmark that assesses both video quality and trajectory control accuracy across different numbers of objects. Extensive experiments demonstrate that MagicMotion outperforms previous methods across various metrics.*

1. Introduction

With the rapid development of diffusion models, video generation has witnessed significant progress in recent years. In contrast to images, videos [53, 63, 64, 73] capture dynamic temporal changes and offer a richer representation of semantic content. Earlier video generation approaches, such as AnimateDiff [16] and SVD [3], primarily rely on the UNet [46] structure, which results in videos with limited length and quality. Sora [4] has demonstrated the powerful capabilities of the DiT [37] architecture in text-to-video (T2V) generation. Following this breakthrough, subsequent models leveraging the DiT architecture [28, 82, 93] have achieved higher quality outputs and extended video duration.

While DiT-based models [28, 82, 93] excel at producing high-quality and longer videos, many text-to-video approaches [16, 75, 77] lack precise control over attributes like object movement and camera motion [16, 75, 82]. Fine-grained trajectory-controllable video generation emerges as a solution, which is especially critical for generating controllable videos in real-world scenarios.

Previous trajectory-controllable video generation methods can be categorized based on the type of control signals they use. These include points-based control [14, 59, 66, 94], optical flow-based control [5, 30, 48, 84, 91], bounding boxes-based control [22, 35, 39, 60, 67, 69, 81], masks-based control [83], and 3D trajectories-based control [13, 15, 62]. However, these methods exhibit several limitations. First, the trajectory control conditions used by these methods are unitary. Each method only support a single type of control signal. However, sparse trajectories (*e.g.*, points and optical flow) result in imprecise control over object shape and size, while dense trajectories (*e.g.*, masks and 3D trajectories) are challenging for users to provide. Second, a publicly available large-scale dataset for trajectory-controllable video generation is lacking. Existing VOS (Video Object Segmentation) datasets suffer from short video lengths [33, 34, 80], small scales [9, 10, 20, 38], or few foreground objects [12]. Third, a unified benchmark for evaluating different methods is absent. Besides, previous work focuses solely on video quality and trajectory accuracy while neglecting the impact of the number of moving objects. We argue that controlling fewer or more objects presents different challenges, making it essential to include this factor in the evaluation metrics.

To address these issues, we propose MagicMotion, a controllable video generation framework with dense-to-sparse trajectory guidance. To inject trajectory conditions into the generation process, we utilize an architecture similar to ControlNet [89] called **Trajectory ControlNet** to encode the trajectory information, which is later added to the original DiT model through a zero-initialized convolution layer. MagicMotion supports three types of trajectory conditions: masks, boxes, and sparse boxes using a progressive training strategy. Experiments show that the model can leverage the knowledge learned in the previous stage to achieve better performance than training from scratch. Additionally, we propose a novel latent segment loss that helps the video generation model better understand the fine-grained shape of objects with minimal computation.

We also construct MagicData, a high-quality public dataset comprising 23K video samples, each annotated with a `<video, text, trajectory>` triplet. We design a data pipeline that uses a large language model [52] to extract the main moving objects in the video, and Segment Anything Model (SAM2) [43, 44] to annotate the segmentation masks and bounding boxes of these objects. Furthermore, we introduce MagicBench, a large-scale comprehensive trajectory-controllable video generation benchmark. It categorizes all videos into 6 classes based on the number of foreground objects and evaluates models separately for each category in terms of both video generation quality and trajectory control accuracy.

In conclusion, the main contributions of our work are summarized as follows:

- We present MagicMotion, a trajectory-controllable image-to-video generation model that supports three types of control signals: masks, boxes, and sparse boxes.
- We introduce a data curation and filtering mechanism, and construct MagicData, the first public dataset for trajectory-controlled video generation.
- We propose MagicBench, a comprehensive benchmark for evaluating trajectory-controllable video generation models for both video quality and trajectory control accuracy across different numbers of controlled objects.

2. Related Works

Video Diffusion Models Diffusion models [18, 31, 49, 50] have made great progress in image generation [8, 36, 40, 41, 45, 47, 61, 85–88, 90], which has led to the rapid development of video generation [3, 11, 16, 19, 21, 25, 28, 51, 54–57, 70, 72, 74, 76, 77, 82, 93]. VDM [19] is the first to apply diffusion models to video generation. Early works like AnimateDiff [16] and SimDA [75] attempt to insert temporal layers into pretrained T2I model for video generation. Subsequently, VideoCrafter [6] and SVD [3] use large-scale and high-quality data for training, acheiving

better performance. However, these methods have trouble on generating long videos with high quality, mainly due to the inherent limitations of the UNet architecture. The emergence of Sora [4] is a significant success, demonstrating the potential of DiT [37] models to generate high-quality videos with tens of seconds. Recent video generation methods [28, 82, 93] are mainly based on DiT architecture, and have achieved great success in the open-source community. However, these methods rely solely on text or image guidance for video generation, lacking precise control over object or camera trajectory, which is crucial for high-quality video generation.

Trajectory Controllable Video Generation Trajectory Controllable Video Generation has recently garnered significant attention for its ability to precisely control object and camera trajectories during video synthesis. Previous methods [30, 48, 84, 91] integrate optical flow maps into video generation through a trajectory encoder. Recent works [66, 94] suggest using point maps as a form of guidance. MotionCtrl [66] processes point maps with a Gaussian filter and employs trainable encoders to encode object trajectories. Trackgo [94] represents objects using a few key points and injects this information via an encoder and a custom-designed adapter structure. Other works [22, 35, 39, 60] employ bounding box to control object trajectories. Boximator [60] employs a trainable self-attention layer to fuse box and visual tokens inspired by GLIGEN [29]. Some training-free methods [22, 35, 39] purpose to modify attention layers or initial noised video latents to inject box signals. Additionally, certain methods [13, 15, 59, 62] explore the potential of 3D trajectories to achieve more sophisticated motion control. LeViTor [59] employs keypoint trajectory maps enriched with depth information, while others [13, 15, 62] construct custom 3D trajectories to represent object movements. However, sparse trajectories lead to imprecise control on objects shape and size, while dense trajectories are difficult for users to provide. In contrast, MagicMotion can control both dense and sparse trajectories, providing users with more flexible control over video generation.

3. Method

3.1. Overview

Our work mainly focuses on trajectory-controllable video generation. Given an input image $I \in R^{H\times W\times 3}$, and several trajectory maps $C \in R^{T\times H\times W\times 3}$, the model can generate a video $V \in R^{T\times H\times W\times 3}$ in line with the provided trajectories, where T denotes the length of generated video. In the following sections, we first provide a detailed explanation of our model architecture in Section 3.2. Next, we outline our progressive training procedure in Section 3.3. In Section 3.4, we introduce the Latent Segmentation Loss and demonstrate how it enhances the model capabilities on fine-grained object shape. We then describe our dataset curation and filtering pipeline in Section 3.5. Finally, we present an in-depth overview of MagicBench in Section 3.6.

3.2. Model Architecture

Base I2V generation model We utilize CogVideoX-5B-I2V [82] and Wan2.1 1.3B [58] as our base image to video model. The models are built upon a DiT (Diffusion Transformer) architecture, incorporating 3D-Full Attention to generate high-quality videos. As shown in Fig. 2, the model takes an input image $I \in R^{H\times W\times 3}$ and a corresponding video $V \in R^{T\times H\times W\times 3}$ and encodes them into latent representations $Z_{image}, Z_{video} \in R^{\frac{T}{4}\times\frac{H}{8}\times\frac{W}{8}\times 16}$ using a pretrained 3D VAE [26]. Later, Z_{image} is zero-padded to T frames and concatenated with a noised version of Z_{video} and then fed into the Diffusion Transformer, where a series of Transformer blocks iteratively denoise it over a predefined number of steps. Finally, the denoised latent is decoded by a 3D VAE decoder to get the output video $V_{out} \in R^{T\times H\times W\times 3}$.

Trajectory ControlNet To ensure that the generated video follows the motion patterns given by the input trajectory maps $C \in R^{T\times H\times W\times 3}$, we adopt a design similar to ControlNet [89] to inject trajectory condition. As shown in Fig. 2, we employ the 3D VAE encoder to encode the trajectory maps into $Z_{trajectory} \in R^{\frac{T}{4}\times\frac{H}{8}\times\frac{W}{8}\times 16}$, which is then concatenated with the encoded video Z_{video} and serves as input to Trajectory ControlNet. Specifically, Trajectory ControlNet is constructed with a trainable copy of all pre-trained DiT blocks to encode the user-provided trajectory information. The output of each Trajectory ControlNet block is then processed through a zero-initialized convolution layer and added to the corresponding DiT block in the base model to provide trajectory guidance.

3.3. Dense-to-Sparse Training Procedure

Dense trajectory conditions, such as segmentation masks, offer more precise control than sparse conditions like bounding boxes but are less user-friendly. To address this, MagicMotion employs a progressive training procedure, where each stage initializes its model with the weights from the previous stage. This enables three types of trajectory control ranging from dense to sparse. We found that this progressive training strategy helps the model achieve better performance compared to training from scratch with sparse conditions. Specifically, we adopt the following trajectory conditions across stages: stage1 uses segmentation masks, stage2 uses bounding boxes, and stage3 uses sparse bounding boxes, where fewer than 10 frames have box annotations. Additionally, we always set the first frame of the trajectory condition as a segmentation mask to specify the foreground objects that should be moving.

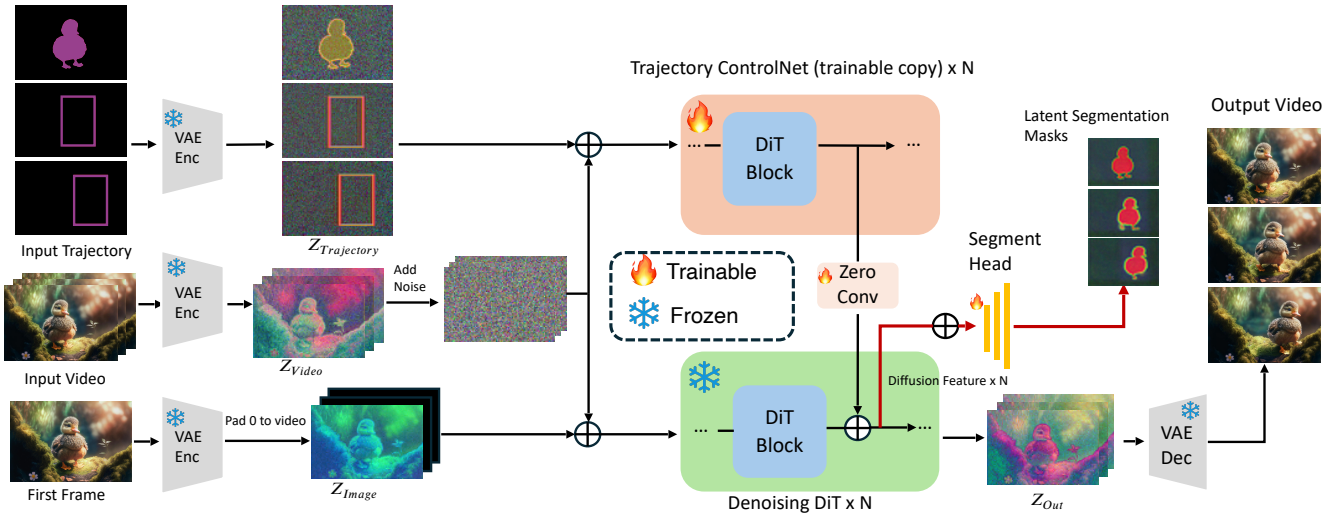


Figure 2. Overview of MagicMotion Architecture (text prompt and encoder are omitted for simplicity). MagicMotion employs a pretrained 3D VAE to encode the input trajectory, first-frame image, and training video into latent space. It has two separate branches: the video branch processes video and image tokens, and the trajectory branch uses Trajectory ControlNet to fuse trajectory and image tokens, which is later integrated to the video branch through a zero-initialized convolution layer. Besides, diffusion features from DiT blocks are concatenated and processed by a trainable segment head to predict latent segmentation masks, which contribute to our latent segment loss.

Our model uses velocity prediction following [82]. Let x_0 be the initial video latents, ϵ be the gaussian noise, $x_t = \sqrt{\alpha_t} * x_0 + \sqrt{1-\alpha_t} * \epsilon$ be the noised video latents, and v_θ be the model output. The diffusion loss can be written as:

$$\mathcal{L}_{diffusion} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I),x_0}\left[\left\|x_0 - \left(\sqrt{\alpha_t}\,x_t - \sqrt{1-\alpha_t}\,v_\theta\right)\right\|_2^2\right] \quad (1)$$

3.4. Latent Segmentation Loss

Bounding box-based trajectory is able to control an object's position and size but lacks fine-grained shape perception. To address this, we propose Latent Segmentation Loss, which introduces segmentation mask information during model training and enhances the model's ability to perceive fine-grained object shapes.

Previous works [2, 68, 79, 92] have leveraged diffusion generation models for perception tasks, demonstrating that the features extracted by diffusion models contain rich semantic information. However, these models generally operate in the pixel space, which leads to extensive computational time and substantial GPU memory.

To incorporate dense trajectory information while keeping computational costs within a reasonable range, we propose utilizing a lightweight segmentation head to predict segmentation masks directly in the latent space, eliminating the need for decoding operations. Specifically, our segmentation head takes a list of diffusion features $Z_{feature} \in R^{\frac{T}{4}*\frac{H}{16}*\frac{W}{16}*3072}$ from each DiT block, and outputs a latent segmentation mask $Z_{segment} \in R^{\frac{T}{4}*\frac{H}{8}*\frac{W}{8}*16}$. We use a light-weight architecture inspired by Panoptic FPN [27]. Each diffusion feature first passes through a convolution layer to extract visual features $P \in R^{\frac{T}{4}*\frac{H}{16}*\frac{W}{16}*64}$. The resulting features are then concatenated and processed by another convolution layer followed by an upsampling layer to generate the final latent segmentation mask. We compute the latent segment loss as the Euclidean distance between $Z_{segment}$ and the ground truth mask trajectory latents Z_{mask}, which can be written as,

$$\mathcal{L}_{seg} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I),z_0}[\|Z_{segment} - Z_{mask}\|_2^2] \quad (2)$$

In practice, $\mathcal{L}_{seg}$ is only used in stage2 and stage3, providing the model with dense conditions information when trained with sparse conditions. In detail, we set the weight of $\mathcal{L}_{seg}$ as 0.5, and the original diffusion loss as 1. In total, our final loss function can be written as,

$$\mathcal{L} = \mathcal{L}_{diffusion} + \lambda * \mathcal{L}_{seg} \quad (3)$$

where λ is set to 0 in stage1, and 0.5 in stage2 and stage3.

3.5. Data Pipeline

Trajectory controllable video generation requires a video dataset with trajectory annotations. However, existing large-scale video datasets [1, 7, 24] only provide text annotations and lack trajectory data. Moreover, almost all previous works [30, 66, 84, 91, 94] use privately curated datasets, which are not publicly available.

We present a comprehensive and general data pipeline for generating high-quality video data with both dense (mask) and sparse (bounding box) annotations. As shown in Fig. 3, the pipeline consists of two main stages: the Curation Pipeline and the Filtering Pipeline. The Curation

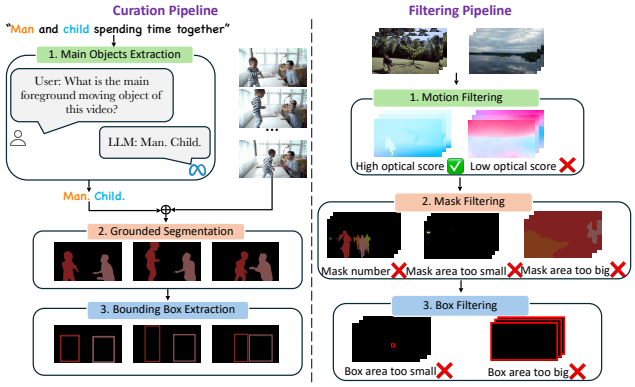


Figure 3. Overview of the Dataset Pipeline. The Curation Pipeline is used to construct trajectory annotations, while the Filtering Pipeline filters out unsuitable videos for training.

Pipeline is responsible for constructing trajectory information from a video-text dataset, while the Filtering Pipeline ensures that unsuitable videos are removed before training.

Curation Pipeline. We begin our dataset curation process with Pexels [23], a large-scale video-text dataset containing 396K video clips with text annotations. It encompasses videos featuring diverse subjects, various scenes, and a wide range of movements. We utilize Llama3.1 [52] to extract the foreground moving objects from the textual annotations of each video. As shown in Fig. 3, we input the video's caption into the language model and prompt it to identify the main foreground objects mentioned in the sentence. If the model determines that the sentence does not contain any foreground objects, it simply returns "empty" and such videos are filtered out. Next, we utilize Grounded-SAM2 [43, 44], a grounded segmentation model that takes a video along with its main objects as input and generates segmentation masks for each primary object. Each object is consistently annotated with a unique color. Finally, bounding boxes are extracted from each segmentation mask using the coordinates of the top-left and bottom-right corners to draw the corresponding boxes. The color of the bounding box for each object remains consistent with its masks.

Filtering Pipeline. Many videos contain only static scenes, which are not beneficial for training trajectory-controlled video generation models. To address this, we use optical flow scores to filter out videos with little motion and dynamics. Specifically, we utilize UniMatch [78] to extract optical flow maps between frames and compute the mean absolute value of these flow maps as the optical flow score, representing the video's motion intensity. However, videos with background movement but static foreground can still have high motion scores. To address this, we further use UniMatch to extract optical flow scores for foreground objects based on segmentation masks and bounding boxes. Videos with low foreground optical flow scores are filtered out, ensuring MagicData includes only videos with moving foreground objects.

The trajectory annotations generated by the curation pipeline require further refinement. As shown in Fig. 3, some videos contain too many foreground objects annotations, or their sizes may be too large or too small. To address this, we regulate these factors within a reasonable range and filter out videos that fall outside the acceptable range.

Specifically, based on extensive manual evaluation, we empirically set the optical flow score threshold to 2.0, limit the number of foreground object annotations from 1 to 3, and constrain the annotated area ratio to a range of 0.008 to 0.83. The whole data curation and filtering pipeline yields us with MagicData, a high quality dataset for trajectory controllable video generation containing 23K videos with both dense and sparse trajectory annotations.

3.6. MagicBench

Previous works on trajectory-controlled video generation [30, 35, 48, 65, 71, 91, 94] have primarily been validated on DAVIS (which has a relatively small dataset size), VIPSeg (where the annotated frames per video are insufficient), or privately constructed test sets. Thus there is an urgent need for a large-scale, publicly available benchmark to enable fair comparisons across different models in this field. To merge this gap, we use the data pipeline mentioned in Sec. 3.5 to construct MagicBench, a large-scale open benchmark consisting of 600 videos with corresponding trajectory annotations. MagicBench evaluates not only video quality and trajectory accuracy but also takes the number of controlled objects as a key evaluation factor. Specifically, it is categorized into 6 groups based on the number of controlled objects, ranging from 1 to 5 objects and more than 5 objects, with each category containing 100 high-quality videos.

Metrics. For evaluation metrics, we adopt FVD [42] to assess video quality and FID [17] to evaluate image quality, following [59, 66, 71, 94]. To quantify motion control accuracy, we use Mask_IoU and Box_IoU, which measure the accuracy of masks and bounding boxes, respectively. Specifically, given a generated video V_{gen}, we use the groundtruth masks of the first frame $M_{gt}(0)$ as input to SAM2 [43] to predict the masks M_{gen} of the foreground objects in V_{gen}. For each foreground object, we compute the Intersection over Union (IoU) between $M_{gen}(i)$ and groundtruth masks $M_{gt}(i)$ in each frame, then average these values to obtain Mask_IoU. Similarly, we compute the IoU between the predicted and groundtruth bounding boxes for each foreground object in each frame and take the average as Box_IoU.

Method	MagicBench				DAVIS			
	FID(↓)	FVD(↓)	M_IoU%(↑)	B_IoU%(↑)	FID(↓)	FVD(↓)	M_IoU%(↑)	B_IoU%(↑)
Motion-I2V [48]	39.31	354.10	56.19	60.66	94.14	1558.32	32.00	42.61
ImageConductor [30]	49.95	331.76	51.76	52.90	91.71	1155.08	34.39	43.41
DragAnything [71]	31.36	253.40	66.30	70.85	70.70	1166.22	40.13	53.60
LeViTor [59]	38.32	194.53	39.96	46.36	97.98	922.68	25.24	31.42
DragNUWA [84]	39.73	185.52	66.88	69.21	84.61	1079.89	41.22	52.61
SG-I2V [35]	32.60	168.82	68.78	74.39	90.93	1170.60	37.36	50.96
Tora [91]	26.27	245.23	58.95	64.03	51.75	766.76	37.98	50.90
Ours (Stage1-Wan1.3B)	**17.74**	**129.79**	**83.79**	**80.96**	**46.65**	**548.56**	**42.83**	**53.67**
Ours (Stage2-Wan1.3B)	**19.33**	**143.07**	**68.89**	**71.74**	**48.33**	**662.82**	**44.75**	**61.88**
Ours (Stage1-CogVideoX)	**15.06**	**112.69**	**91.57**	**87.75**	**45.06**	**579.94**	**81.33**	**84.97**
Ours (Stage2-CogVideoX)	**15.17**	**107.21**	**76.61**	**81.45**	**50.36**	**760.95**	**53.94**	**72.84**

Table 1. Quantitative Comparison results on MagicBench and DAVIS. "M_IoU" and "B_IoU" refer to Mask_IoU and Box_IoU, respectively.

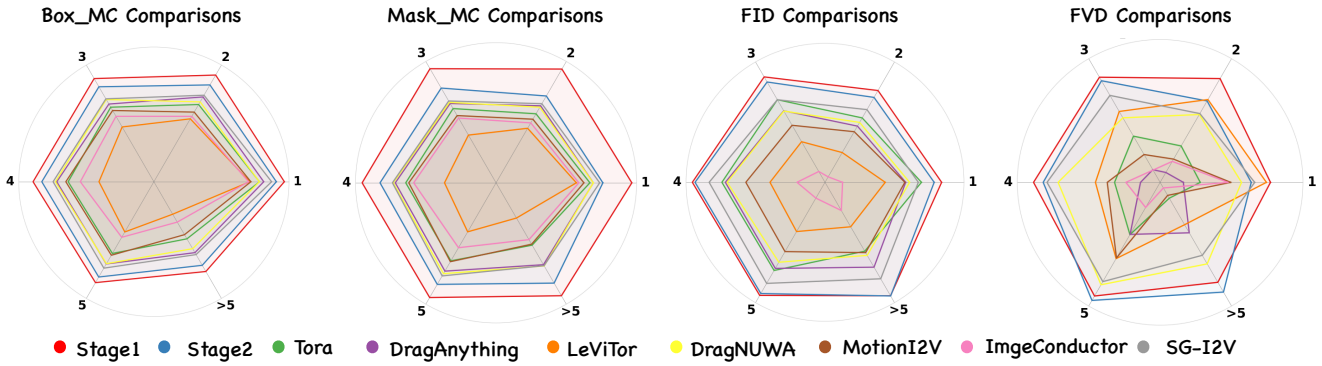


Figure 4. Comparison results of different object number on MagicBench. To present the results more clearly, we have negated the FVD and FID scores.

4. Experiment

4.1. Experiment Settings

Implementation details. We employ CogVideoX 5B [82] and Wan2.1 1.3B [58] as our base image-to-video model, which is trained to generate video based on an input image. Each stage of MagicMotion was trained on MagicData for one epoch. The training process consists of three stages. stage1 trains Trajectory ControlNet from scratch. In stage2, Trajectory ControlNet is further refined using the weights from stage1, while Segment Head is trained from scratch. Finally, in stage3, both Trajectory ControlNet and Segment Head continue training initialized with the weights from stage2. All training experiments were conducted on 4 NVIDIA A100-80G GPUs. We employed AdamW [32] as the optimizer, training with a learning rate of $1e-5$ and a batch size of 1 on each GPU. During inference, we set steps to 50, the guidance scale to 6, and the weight of Trajectory ControlNet to 1.0 by default.

Datasets. During training, we use MagicData as our training set. MagicData is annotated with dense to sparse trajectory information using the data pipeline described in Section 3.5. It comprises a total of 23K video samples, annotated with text and trajectory. For evaluation, we evaluate all methods on both MagicBench and DAVIS [38], using the comparison metrics illustrated in Sec 3.6.

4.2. Comparison with Other Approaches

For thorough and fair comparisons, we compare our methods against 7 public trajectory controllable I2V methods [30, 35, 48, 59, 71, 84, 91]. Quantitative comparison and qualitative comparison results are shown below.

Quantitative comparison. To compare MagicMotion with previous works, we use the first 49 frames of each video from DAVIS and MagicBench as the ground truth video. Since some methods [30, 35, 48, 59, 71, 84] do not support video generation up to 49 frames in length, we uniformly sample N frames from these 49 frames for evaluation, where N represents the video length that each method support. We leverage the mask and box annotations from these selected frames as trajectory inputs for mask or box-based methods. The center point of each frame's

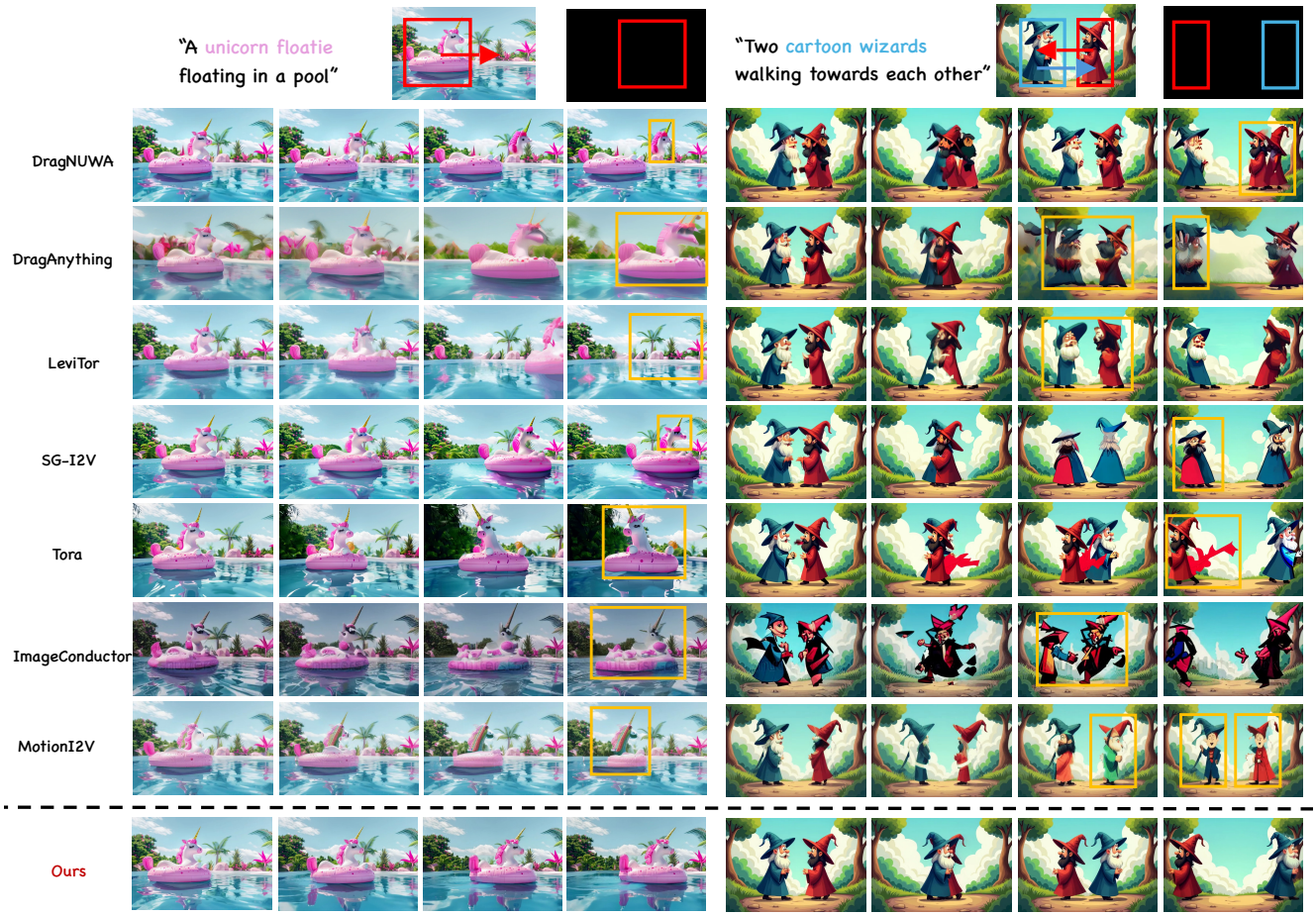


Figure 5. MagicMotion successfully controls the main objects moving along the provided trajectory, while all other methods exhibit significant defects marked with the orange box.

mask is extracted as input for point or flow-based methods [30, 48, 59, 71, 84, 91].

As shown in Table 1, our method outperforms all previous approaches across all metrics both on MagicBench and DAVIS, demonstrating its ability to produce higher-quality videos and more precise trajectory control. Additionally, we evaluate each method's performance on MagicBench based on the number of controlled objects. As shown in Fig. 4, our method achieves the best results across all object number categories, demonstrating the superiority of our approach.

Qualitative comparison. Qualitative comparison results are shown in Fig. 5, with the input image, prompt and trajectory provided. As shown in Fig. 5, Tora [91] accurately controls the motion trajectory but struggles to maintain the shape of the objects. While DragAnything [71], ImageConductor [30], and MotionI2V [48] have difficulty preserving the consistency of the original subject, resulting in substantial deformation in subsequent frames. Meanwhile, DragNUWA [84], LeviTor [59], and SG-I2V [35] frequently produce artifacts and inconsistencies in fine details. In contrast, MagicMotion allows moving objects to follow the specified trajectory smoothly while preserving high video quality.

4.3. Ablation Studies

In this section, we present ablation studies to validate the effectiveness of our MagicData dataset. Additionally, we demonstrate how our progressive training procedure and latent segment loss enhance the model's understanding of precise object shapes under sparse control conditions, thereby improving trajectory control accuracy.

Ablations on Dataset. To verify the effectiveness of MagicData, we constructed an ablation dataset by combining two public VOS datasets, MeViS [9] and MOSE [10]. For a fair comparison, we trained MagicMotion stage2 for one epoch using either MagicData or the ablation dataset as the training set, both initialized with the same stage1 weights. We then evaluated the models on both MagicBench and DAVIS.

As shown in Table 2, the model trained on MagicData

outperforms the one trained on the ablation dataset across all metrics. Qualitative comparisons are shown in Fig. 6. In this case, we aim to gradually move the boy in the lower right corner to the center of the image. However, not using MagicData results to an unexpected child appears next to the boy. In contrast, the model trained with MagicData performs well, moving the boy along the specified trajectory while maintaining video quality.

Method	MagicBench/DAVIS			
	FID(↓)	FVD(↓)	M_IoU%(↑)	B_IoU%(↑)
w/o MagicData	24.24/50.91	125.19/768.27	73.39/50.95	78.66/70.05
Ours	**15.17/50.35**	**107.21/760.95**	**76.61/53.94**	**81.45/72.84**

Table 2. Ablation Study on MagicData. The model trained with MagicData outperforms the one trained without it across all metrics.

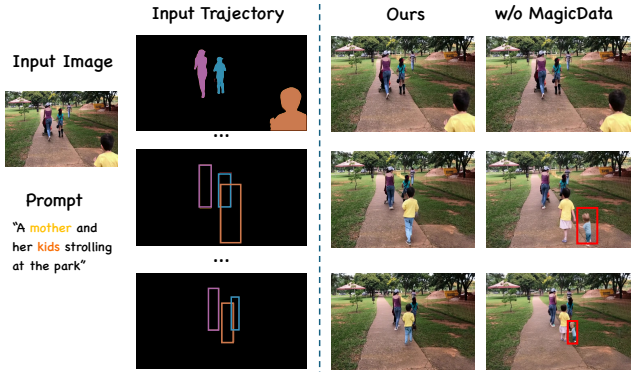


Figure 6. Ablation Study on MagicData. Not using MagicData causes the model to generate an unexpected child.

Ablations on Progressive Training Procedure. Progressive Training Procedure allows the model to leverage the weights learned in the previous stage, incorporating dense trajectory control information when trained with sparse trajectory conditions. To validate the effectiveness of this approach, we train the model from scratch for one epoch using bounding boxes as trajectory conditions. We then compare its performance with MagicMotion stage2.

As shown in Table 3, excluding Progressive Training Procedure weakens the model's ability to perceive object shapes, ultimately reducing the accuracy of trajectory control. Qualitative comparisons in Fig. 7 further illustrate these effects, where the model trained without Progressive Training Procedure turns the woman's head entirely into hair.

Ablations on Latent Segment Loss. Latent Segment Loss makes the model predict dense segmentation masks while training with sparse trajectories, enhancing its ability to perceive fine-grained object shapes under sparse conditions. To evaluate the effectiveness of this technique, we train the model from stage1 for one epoch using bounding boxes as trajectory conditions and compare its performance with MagicMotion stage2. Table 3 shows that the absence of Latent Segment Loss reduces the model's ability on object shapes, leading to less precise trajectory control.

Method	MagicBench		DAVIS	
	M_IoU%(↑)	B_IoU%(↑)	M_IoU%(↑)	B_IoU%(↑)
w/o LSL	74.65	78.02	49.19	64.19
w/o PT	74.61	79.49	49.22	67.34
Ours	**76.61**	**81.45**	**53.94**	**72.84**

Table 3. Ablation Study on Progressive Training (PT) and Latent Segment Loss (LSL). Experimental results demonstrate that these techniques enhance the model with better comprehension on fine-grained object shapes.

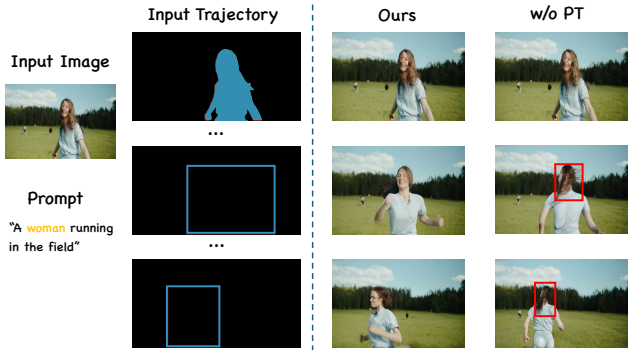


Figure 7. Ablation study on the progressive training procedure. Without it, the generated head shapes become noticeably distorted.

5. Conclusion

In this paper, we proposed MagicMotion, a trajectory-controlled image-to-video generation method that uses a ControlNet-like architecture to integrate trajectory information into the diffusion transformer. We employed a progressive training strategy, allowing MagicMotion to support three levels of trajectory control: dense masks, bounding boxes and sparse boxes. We also utilized Latent Segment Loss to enhance the ability of the model to perceive fine-grained object shapes when only provided with sparse trajectory conditions. Additionally, we presented MagicData, a high-quality annotated dataset for trajectory-controlled video generation, created through a robust data pipeline. Finally, we introduced MagicBench, a large-scale benchmark for evaluating trajectory-controlled video generation. MagicBench not only assessed video quality and trajectory accuracy but also took the number of controlled objects into account. Extensive experiments on both MagicBench and DAVIS demonstrated the superiority of MagicMotion compared to previous works.

Acknowledge This work was supported in part by the National Natural Science Foundation of China (Grant 62032006 and Grant 62472098).

References

[1] 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

[2] Aditya Bhat, Rupak Bose, Chinedu Innocent Nwoye, and Nicolas Padoy. Simgen: A diffusion-based framework for simultaneous surgical image and segmentation mask generation. *arXiv preprint arXiv:2501.09008*, 2025. 4

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

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *https://openai.com/research/video-generation-models-as-world-simulators*, 2024. 2, 3

[5] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, 2025. 2

[6] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[7] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 4

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

[9] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 2, 7

[10] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 2, 7

[11] Zhiying Du and Zhen Xing. Challenge report: Track 2 of multimodal perception and comprehension of corner cases in autonomous driving. In *ECCV Workshop*, 2024. 2

[12] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *IJCV*, 2021. 2

[13] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In *ICLR*, 2025. 2, 3

[14] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 2

[15] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 2, 3

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ICLR*, 2024. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 5

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

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 35, 2022. 2

[20] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *ICCV*, 2023. 2

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

[22] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. *arXiv preprint arXiv:2312.07509*, 2023. 2, 3

[23] JovianZM. Pexels-400k dataset. https://huggingface.co/datasets/jovianzm/Pexels-400k, 2024. 5

[24] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *NeurIPS*, 2024. 4

[25] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[27] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 4

[28] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3

[29] 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. *CVPR*, 2023. 3

[30] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *AAAI*, 2025. 2, 3, 4, 5, 6, 7

[31] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[33] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021. 2

[34] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 2

[35] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. In *ICLR*, 2025. 2, 3, 5, 6, 7

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

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2, 3

[38] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 6

[39] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024. 2, 3

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

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

[42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 5

[43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 5

[44] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 5

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

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

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

[48] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *SIGGRAPH*, 2024. 2, 3, 5, 6, 7

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

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

[51] Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025. 2

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 5

[53] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *ICCV*, 2023. 2

[54] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024. 2

[55] Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024.

[56] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *CVPR*, 2025.

[57] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*, 2025. 2

[58] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 6

[59] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. Levitor: 3d trajectory oriented image-to-video synthesis. In *CVPR*, pages 12490–12500, 2025. 2, 3, 5, 6, 7

[60] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 2, 3

[61] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025. 2

[62] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *SIGGRAPH*, 2025. 2, 3

[63] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022. 2

[64] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023. 2

[65] Zhouxia Wang, Yushi Lan, Shangchen Zhou, and Chen Change Loy. ObjCtrl-2.5D: Training-free object control with camera poses. In *arXiv preprint arXiv:2412.07721*, 2024. 5

[66] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024. 2, 3, 4, 5

[67] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024. 2

[68] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Genrec: Unifying video generation and recognition with diffusion models. *NeurIPS*, 2024. 4

[69] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *NeurIPS*, 2024. 2

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

[71] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *ECCV*. Springer, 2024. 5, 6, 7

[72] Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Shiwei Shen, Jiaqi Leng, Xipeng Qiu, Yanwei Fu, Zuxuan Wu, and Yu-Gang Jiang. Human2robot: Learning robot actions from paired human-robot videos. *arXiv preprint arXiv:2502.16587*, 2025. 2

[73] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *CVPR*, 2023. 2

[74] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023. 2

[75] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *CVPR*, 2024. 2

[76] Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. *arXiv preprint arXiv:2406.06465*, 2024. 2

[77] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2024. 2

[78] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 5

[79] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.04803*, 2023. 4

[80] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[81] 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 *SIGGRAPH*, 2024. 2

[82] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video

diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3, 4, 6

[83] Guy Yariv, Yuval Kirstain, Amit Zohar, Shelly Sheynin, Yaniv Taigman, Yossi Adi, Sagie Benaim, and Adam Polyak. Through-the-mask: Mask-based motion trajectories for image-to-video generation. In *CVPR*, 2025. 2

[84] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2, 3, 4, 6, 7

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

[86] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. *arXiv preprint arXiv:2412.03859*, 2024.

[87] Hui Zhang, Tingwei Gao, Jie Shao, and Zuxuan Wu. Blockdance: Reuse structurally similar spatio-temporal features to accelerate diffusion transformers. In *CVPR*, 2025.

[88] Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Chen, Zhao Zhang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. *arXiv preprint arXiv:2505.19114*, 2025. 2

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

[90] Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl. *arXiv preprint arXiv:2505.24875*, 2025. 2

[91] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7

[92] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023. 4

[93] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2, 3

[94] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: A flexible and efficient method for controllable video generation. In *AAAI*, 2025. 2, 3, 4, 5