Mask2IV: Interaction-Centric Video Generation via Mask Trajectories

Gen Li[1,3], **Bo Zhao**[2*], **Jianfei Yang**[1*], **Laura Sevilla-Lara**[3]
[1]Nanyang Technological University
[2]Shanghai Jiao Tong University
[3]University of Edinburgh
{gen.li, jianfei.yang}@ntu.ed.sg, bo.zhao@sjtu.edu.cn

Abstract

Generating interaction-centric videos, such as those depicting humans or robots interacting with objects, is crucial for embodied intelligence, as they provide rich and diverse visual priors for robot learning, manipulation policy training, and affordance reasoning. However, existing methods often struggle to model such complex and dynamic interactions. While recent studies show that masks can serve as effective control signals and enhance generation quality, obtaining dense and precise mask annotations remains a major challenge for real-world use. To overcome this limitation, we introduce Mask2IV, a novel framework specifically designed for interaction-centric video generation. It adopts a decoupled two-stage pipeline that first predicts plausible motion trajectories for both actor and object, then generates a video conditioned on these trajectories. This design eliminates the need for dense mask inputs from users while preserving the flexibility to manipulate the interaction process. Furthermore, Mask2IV supports versatile and intuitive control, allowing users to specify the target object of interaction and guide the motion trajectory through action descriptions or spatial position cues. To support systematic training and evaluation, we curate two benchmarks covering diverse action and object categories across both human-object interaction and robotic manipulation scenarios. Extensive experiments demonstrate that our method achieves superior visual realism and controllability compared to existing baselines.

Project page — https://reagan1311.github.io/mask2iv

Introduction

Interacting with objects is a fundamental aspect of daily human life: we actively engage with a wide variety of objects without explicit prior planning. Our extensive experience with embodied interaction allows us to anticipate the appropriate hand pose, the trajectory required for a specific action, and the ideal placement of a target—even before physical contact occurs. This remarkable imaginative ability stems from our rich prior knowledge of object dynamics and action understanding. However, despite appearing effortless for humans, replicating this capacity in generative models or robotic systems remains a formidable challenge. In the context of embodied AI, addressing this challenge is particularly valuable, as the ability to synthesize realistic, diverse, and controllable human-object or robot-object interaction sequences can provide powerful visual priors that facilitate tasks such as imitation learning (Kareer et al. 2024; Lepert, Fang, and Bohg 2025) and affordance learning (Li et al. 2025, 2023a). Moreover, high-fidelity modeling of these interactions is crucial for a wide range of applications, including augmented and virtual reality, robot learning, and motion planning.

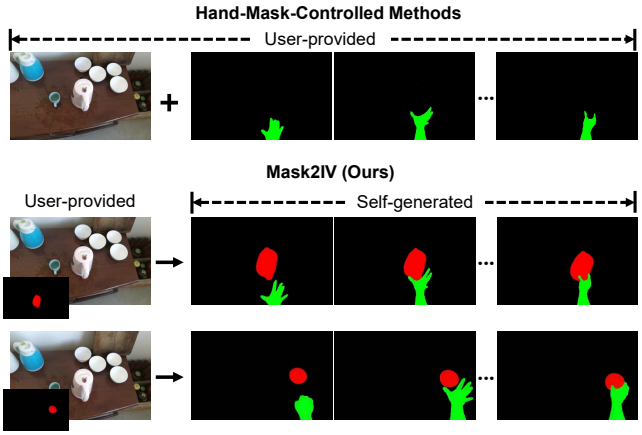

Figure 1: Comparison on control signal acquisition. Hand-mask-controlled video generation methods require users to provide dense hand mask sequences as input. In contrast, Mask2IV autonomously generates trajectories for both hands and objects without manual annotation, and can adaptively produce different trajectories based on the specified object.

Recent advances in diffusion models have substantially improved the generation of realistic and coherent visual content. Nonetheless, current models often fall short in accurately capturing the complex and dynamic interactions between human hands or robotic manipulators and objects. To this end, a growing body of research has emerged, with a focus on interaction-centric image and video generation. Much of the existing work (Lai et al. 2024; Li, Cao, and Corso 2024; Wang et al. 2024b) targets egocentric views, where

*Corresponding authors.

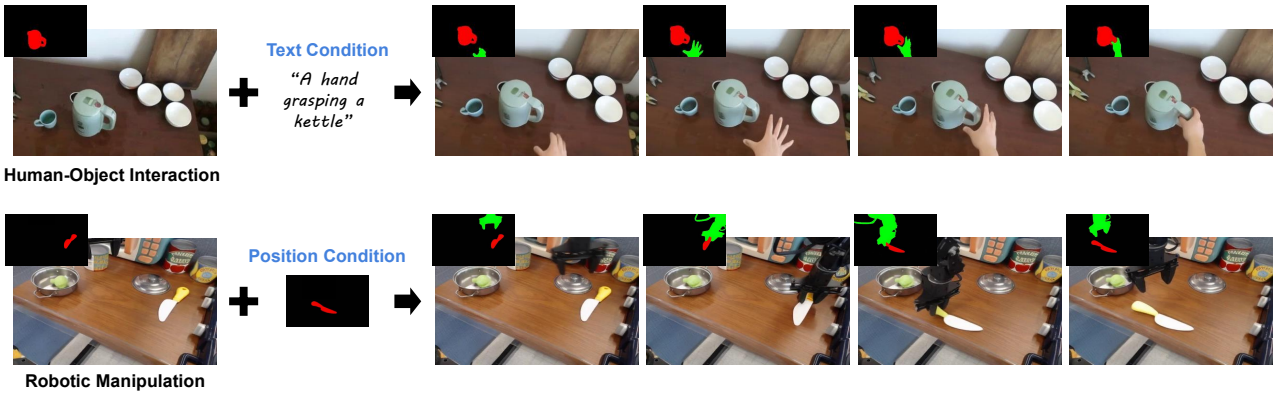

Figure 2: Mask2IV synthesizes videos of human hands or robot arms interacting with a specified object, indicated by an input mask. It first predicts a mask-based interaction trajectory (visualized in the top-left corner of each frame), and then generates the video guided by this trajectory. The generation process is conditioned on either a text prompt or a target position mask.

diverse human-object interaction behaviors are prevalent. While promising results have been achieved, these methods typically rely on text-based conditioning, which is insufficient for modeling fine-grained interactions. Specifically, they lack precise control over which object is being interacted with and where the hands are positioned. To address this shortcoming, recent studies (Sudhakar et al. 2024; Akkerman et al. 2025) have proposed using hand masks as an explicit control signal for interaction modeling. Given an input image of a hand interacting with an object, dense temporal hand masks are provided to guide the generation of subsequent frames, ensuring alignment with the intended interaction. This hand mask control mechanism, while effective for generating plausible interactions, suffers from two critical limitations. First, hand-mask conditioning is inherently impractical and user-unfriendly, as it necessitates frame-by-frame acquisition of precise hand masks. Such masks can only be obtained by first recording or synthesizing the exact interaction video that one wishes to control. In practice, it is unrealistic to expect users to provide a complete sequence of temporal masks in advance. Second, relying exclusively on hand masks constrains the scope of interaction modeling, which hinders both the precise specification of target objects for interaction and the accurate capture of fine-grained hand-object contact, particularly under camera motion.

To overcome these limitations, we propose Mask2IV (Mask-to-Interaction Video generation), a two-stage framework that decomposes the video generation process into interaction trajectory generation and trajectory-conditioned video generation. As illustrated in Fig. 1, unlike hand-mask-controlled methods such as InterDyn (Akkerman et al. 2025), our approach eliminates the need to provide dense mask sequences as control signals. Instead, it predicts interaction trajectories based on an initial image and an object mask. Moreover, Mask2IV produces trajectories of both actors and objects for control, which exhibits improved performance, especially in challenging egocentric and multi-object settings. In addition, Mask2IV supports object-specific control by generating interaction trajectories conditioned on arbitrary target objects.

We demonstrate our model's capabilities across two key domains: human-object interaction and robotic manipulation. To guide the generation of interaction trajectories, we explore two types of conditioning prompts, which are action descriptions and spatial position cues. The former provides a high-level and intuitive specification of the intended action, while the latter enables precise and low-level control over object placement. Examples generated by our approach are presented in Fig. 2, illustrating highly plausible and physically consistent interaction dynamics. For training and evaluation, we curate dedicated benchmarks building on two widely used datasets: HOI4D (Liu et al. 2022) for human-object interaction and BridgeDataV2 (Walke, Black et al. 2023) for robotic manipulation.

To summarize, our main contributions are as follows:

- We introduce the task of interaction-centric video generation, targeting realistic and controllable synthesis of human-object and robot-object interactions.
- We propose an innovative decoupled two-stage framework that not only improves the generation quality but also enables explicit control over the interaction target.
- We construct two dedicated benchmarks covering diverse interaction scenarios and conduct extensive experiments, demonstrating that our approach outperforms existing baselines in both controllability and visual fidelity.

Related Work

Human and Robot Interaction Synthesis. Synthesizing realistic and plausible human or robot interactions is a fundamental problem in computer vision and robotics. A large body of work focuses on generating human interactions in 2D images, with different focuses on hand (Hu et al. 2022; Ye et al. 2023; Narasimhaswamy et al. 2024; Qin et al. 2024; Park, Kong, and Kang 2024; Sudhakar et al. 2024; Lai et al. 2024; Souček et al. 2024) or the whole-body interaction (Kulal et al. 2023; Yang et al. 2024; Hoe et al. 2024; Xu et al. 2024a; Jiang-Lin et al. 2024; Fang et al. 2024).

With the recent progress in video generation, growing attention has turned toward modeling human-object interactions in dynamic scenes (Furuta et al. 2024; Xu et al. 2024b). EgoVid (Wang et al. 2024b) addresses the data bottleneck by introducing a high-quality dataset curated for egocentric video generation. HOI-Swap (Xue et al. 2024) and ReHOLD (Fan et al. 2025) synthesize novel interactions by recombining hands and objects in videos. InterDyn (Akkerman et al. 2025) proposes a controllable generation framework using hand mask sequences, but requires dense annotations and is limited to close-up scenes. In parallel, there has been increasing interest in robot interaction synthesis, particularly in the context of robot-object manipulation. Recent studies such as RoboGen (Wang et al. 2024c), Genesis (Authors 2024), and TesserAct (Zhen et al. 2025) leverage generative models to produce realistic and diverse robot interactions. These methods often target downstream applications such as training visuomotor policies or generating synthetic demonstrations. In this work, we introduce the task of interaction-centric video generation and explore the synthesis of both human and robot interactions within a single framework.

Controllable Video Generation. With the advent of diffusion models (Ho, Jain, and Abbeel 2020), video generation has seen significant advancements, enabling the synthesis of high-quality videos from text-based or image-based prompts. One particular area of interest is controllable video generation, which aims to provide users with the ability to influence the generated content. The success of conditional text-to-image generation methods, such as ControlNet (Zhang, Rao, and Agrawala 2023), GLIGEN (Li et al. 2023b), or T2I-Adapter (Mou et al. 2024), have laid the groundwork for controllable video synthesis. Building on these advances, recent work has extended similar control mechanisms to the video generation. A variety of control signals have been explored, including bounding-boxes (Wang et al. 2024a; Luo et al. 2024), masks (Dai et al. 2023; Yariv et al. 2025; Akkerman et al. 2025), depth maps (Chen et al. 2023; Liang et al. 2024), and optical flow (Shi et al. 2024; Liang et al. 2024). Other efforts focus on camera-level constraints, using camera trajectories or point tracks to guide dynamic viewpoint changes (Wang et al. 2024d; He et al. 2024; Geng, Herrmann et al. 2024). While dense control signals have enabled fine-grained control over layout and motion, they are often impractical and not user-friendly. Providing dense, temporally consistent annotations is especially challenging for dynamic interactions involving actors and objects. To address this, we propose a two-stage approach that first predicts a sequence of masks depicting interaction trajectories, then uses these mask-based trajectories to guide subsequent video generation. This design allows users to automatically obtain dense control signals, while retaining flexibility in controlling generated videos.

Method

We focus on the problem of interaction-centric video generation, with the goal of enabling fine-grained control over the interaction process. Specifically, we aim to satisfy the following three objectives: (1) **Object Specification**. The object involved in the interaction can be explicitly designated. (2) **Action Control**. The interaction can be guided using language descriptions of the intended action. (3) **Target Localization**. The final position of the interacted object can be precisely controlled.

To achieve these goals, we propose Mask2IV that decomposes the generation process into two stages as illustrated in Fig. 3:

- **Interaction Trajectory Generation.** We simplify the problem by first generating the intermediate interaction trajectory, allowing the model to focus solely on motion dynamics without the complexity of appearance details.
- **Trajectory-conditioned Video Generation.** Leveraging the generated trajectory, we synthesize the final video while introducing specialized components to tackle the unique challenges of interaction generation.

This decomposition offers two key advantages: it mitigates the difficulty of directly generating intricate interaction dynamics, and provides greater flexibility for controlling the interaction process. In this section, we first formalize the problem definition, then describe our two-stage framework in detail, and finally present the benchmark construction process for training and evaluation.

Problem Formulation

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and a mask $M \in \mathbb{R}^{H\times W}$ that specifies the object of interest, the target is to generate a video $V \in \mathbb{R}^{N\times H\times W\times 3}$ consisting of N frames that realistically portray dynamic interactions between the designated object and an actor (*e.g.*, human hands or a robot arm). To enable fine-grained control, we incorporate two complementary conditional inputs: (1) a text prompt T that describes the intended action (*e.g.*, a hand picking up a mug), and (2) a target position mask $P \in \mathbb{R}^{H\times W}$ that specifies the desired post-interaction placement of the object. Our model flexibly adapts to either condition, ensuring that the generated video aligns with the described action or reflects the specified target location. In contrast to prior work (Akkerman et al. 2025), our method eliminates the need for densely annotated actor masks during inference, greatly improving the practicability.

Interaction Trajectory Generation

Mask2IV tackles the interaction-centric video generation task with a two-stage framework. In the first stage, the model f_θ is trained to generate an interaction trajectory between the actor and the object, represented as a sequence of masks. It takes as input the initial frame I, the object mask M, and a conditioning signal, either a text prompt T or a position mask P, and outputs the interaction trajectory $S \in \mathbb{R}^{N\times H\times W\times 3}$. Specifically, the input frame I and object mask M are encoded by the VAE encoder $\mathcal{E}$ into latent features $f_i \in \mathbb{R}^{h\times w\times 4}$ and $f_m \in \mathbb{R}^{h\times w\times 4}$, where $h = H/8$ and $w = W/8$. Since $\mathcal{E}$ requires a three-channel input, we first apply color encoding to the object mask M to convert it into the RGB format. If the initial frame contains an actor, we use GroundedSAM (Ren, Liu et al. 2024) to per-

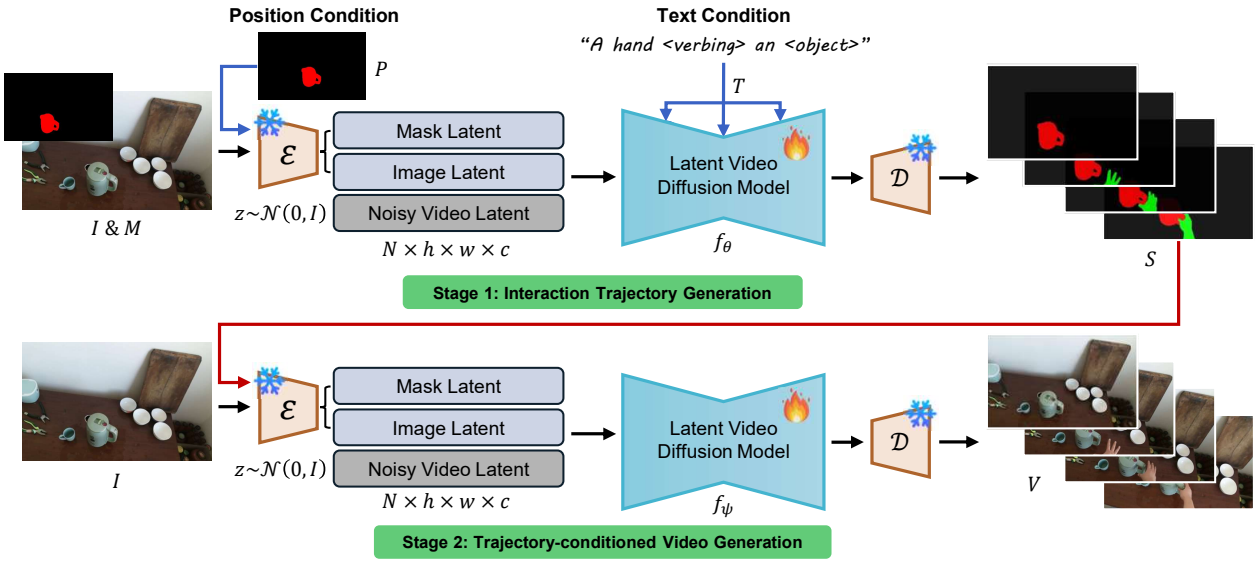

Figure 3: The framework of Mask2IV. It consists of two stages: Interaction Trajectory Generation and Trajectory-conditioned Video Generation. The first stage produces a mask-based interaction trajectory, while the second stage synthesizes a video conditioned on the predicted trajectory.

form segmentation and assign the actor a distinct color from that of the object, enabling the model to better differentiate their roles. The resulting latent features are then combined along the channel dimension, duplicated N times to match the video length, and concatenated with the noise latent $z \in \mathbb{R}^{N \times h \times w \times 4}$. These combined features are fed into the latent video diffusion model, which, in conjunction with the VAE decoder $\mathcal{D}$, generates the interaction trajectory over time. To preserve the motion priors, we freeze the temporal attention layer and fine-tune the remaining parameters of a pre-trained image-to-video diffusion model.

Under this setup, we explore two trajectory generation variants based on different types of conditioning: Text-conditioned Trajectory Generation (TT-Gen) and Position-conditioned Trajectory Generation (PT-Gen). (1) TT-Gen leverages language guidance to shape the trajectory. It enables the model to distinguish between subtle interaction intents, such as pick up vs. put down, or push vs. pull. The text prompt T is encoded using CLIP (Radford et al. 2021) and injected into the model via cross-attention. (2) PT-Gen focuses on precise spatial control of the object's final position. The target position mask P is encoded into a latent feature and inserted in the final frame's slot, while the initial object mask latent is assigned to the first frame. All intermediate frames are filled with zero values, prompting the model to interpolate a coherent trajectory that transitions the object to the specified end position.

Trajectory-conditioned Video Generation

In the second stage, we fine-tune another model f_ψ to generate the interaction video V, which is conditioned on the input image I and the mask-based trajectory S produced in the first stage. Concretely, S is firstly transformed into a latent feature tensor $f_s \in \mathbb{R}^{N \times h \times w \times 4}$ using the VAE encoder $\mathcal{E}$. To enable control, f_s is concatenated with the noise latent z and the first-frame latent feature f_i, which is expanded N times in the temporal dimension to match the video length. The model is then trained with ground-truth mask sequences and infers using predicted masks from the first stage.

While the latent concatenation allows the generated video to follow the trajectory signal, the model lacks specific designs to handle the complex and dynamic interaction behaviors. In particular, we observe two common issues: (1) The synthesized videos are sensitive to variations in the mask trajectory. (2) Regions around the boundary of the actor and the object, i.e., the contact area, is difficult to synthesize accurately. We thus propose two designs to address these challenges. First, we apply random perturbations to the trajectory S to enhance the robustness of the generation process. During training, with a probability p (set to 0.2 in our case), we dilate or erode the mask using a kernel size randomly chosen from $\{3, 5, 7\}$. Since different actors and objects may have masks of varying shapes, this operation encourages the model to generalize better, rather than strictly adhere to the exact shape of the input trajectory. Second, we introduce a contact weighting loss to emphasize content in the interaction-rich regions. Specifically, based on the hand mask m_h and object mask m_o, a contact map m_c is defined as:

$$m_c = (\delta(m_h) \cap m_o) \cup (m_h \cap \delta(m_o)), \quad (1)$$

where $\delta(\cdot)$ denotes the dilation operation. The contact map is then used to reweight the diffusion objective, prioritizing contact regions:

$$w = (1 - m_c) + \lambda \cdot m_c, \quad (2)$$

$$\mathcal{L} = \mathbb{E}_{z,S,\epsilon,t}\left[\|w \odot (\epsilon - \epsilon_\theta(z, f_\psi(S), t))\|_2^2\right], \quad (3)$$

where λ is a weighting factor denoting the importance of

contact regions in the loss, $\epsilon_\theta(\cdot)$ represents the denoising network, and t is the timestep in the forward process.

Benchmark Construction

To align with our focus on interaction-centric video generation, we target two representative and widely studied scenarios: Human-Object Interaction (HOI) and Robotic Manipulation. Both scenarios require the modeling of fine-grained, temporally grounded interactions between actors and objects. We explore text-conditioned trajectory generation for the HOI data, as language provides an intuitive and flexible way to specify human actions. For robotic data, we adopt position-conditioned trajectory generation, since robotics tasks often involve pick-and-place actions that demand precise control over object placement.

We curate benchmarks with frame-level segmentation maps to support model training and evaluation. For HOI, we select the HOI4D (Liu et al. 2022) dataset. It provides timestamped annotations for action start and end points, segmentation masks, and rich hand-object interactions, especially the grasp of diverse objects. We crop the video clip using the timestamps and employ a text prompt template with the form "a hand {verbing} an {object}". To ensure sufficient interaction intensity, we compute a motion score based on the displacement of the hand and the object across each clip. Videos falling below the 5th percentile of this score distribution are filtered out to remove clips with minimal motion. For robotic interactions, we adopt BridgeData V2 (Walke, Black et al. 2023), a large dataset that captures robot manipulation across different environments with variation in objects, camera poses, and workspace configurations. However, it does not provide segmentation masks. We thus utilize GroundingDINO (Liu et al. 2024) for object detection and SAM2 (Ravi, Gabeur et al. 2024) for video segmentation, targeting both the robot arm and the interacting object. Since the data is captured in cluttered scenes with multiple similar objects, we need to further identify the specific object involved in the interaction. To this end, we compute the mean Intersection-over-Union (mIoU) of object masks across temporally spaced frames. Objects with low temporal mIoU are identified as those being manipulated, due to their changing shape and position over time.

Experiments

Experimental Setup

Implementation Details. We build our method on DynamiCrafter (Xing et al. 2024), one of the state-of-the-art image-to-video generation methods, and incorporate extra convolution channels to support encoding of the mask latent. All experiments are conducted on two 80GB NVIDIA A100 GPUs. Each video consists of 16 frames, which are resized and cropped to a resolution of 320×512. We employ AdamW (Loshchilov and Hutter 2019) optimizer with a learning rate of 1×10^{-5} and a batch size of 8. The weighting factor λ is empirically set to 5 during training. At inference, the DDIM sampler (Song, Meng, and Ermon 2020) is used with 50 timesteps to progressively denoise the latent representation and produce the output video.

Metrics. We evaluate the generated videos from two key perspectives: generation quality and text-video alignment. For generation quality, we employ Fréchet Video Distance (FVD) (Unterthiner et al. 2019) to compute overall spatio-temporal similarity of videos, and use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al. 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) to measure the frame-level visual fidelity. To measure text-video alignment, we use pretrained video-language models tailored to each dataset: EgoVLP (Lin et al. 2022) for the HOI4D and ViCLIP (Wang et al. 2023) for the BridgeData V2. We utilize the Text-to-Video Similarity (T2V-Sim) score to measure how closely the generated video matches the text prompt, computed as the cosine similarity between text and video embeddings. Additionally, we compute the Video-to-Video Similarity (V2V-Sim) score, which is used to evaluate the alignment between generated and real videos in the embedding space.

Baselines. We take two types of baseline methods for comparison. The first is the pre-trained image-to-video diffusion model DynamiCrafter, which generates videos from a single input image and a text prompt, but lacks explicit control over the generated interaction. We also include a fine-tuned version (DynamiCrafter-ft) to adapt the model to the curated datasets. The second type consists of control-based method designed for interaction data generation, including CosHand and InterDyn. However, these approaches require ground-truth control signals, which are not available in our setting and thus not directly comparable. To ensure a fair and meaningful comparison, we adopt the same pseudo control signals produced by the first stage of our pipeline. We also adapt their architectures where necessary to align with our setting. Specifically, CosHand was originally proposed for image generation by concatenating the future mask latent with the noise latent. To extend it for video generation, we concatenate mask latent features from all frames to provide temporal conditioning. InterDyn is implemented similarly to its original form, except that we replace the ground-truth actor mask trajectories with our pseudo ones.

Quantitative Analysis

Quantitative comparisons with baselines on HOI4D and BridgeData V2 are shown in Tab. 1. We observe that the original pre-trained DynamiCrafter performs poorly on both datasets, with particularly degraded results on BridgeData V2. This highlights the limitations of current image-to-video generation models in capturing complex human-object or robot-object interactions. After fine-tuning on each dataset, we see a notable improvement in performance, suggesting that domain adaptation helps the model better align with interaction-specific data distributions. However, it lacks the ability to control the motion of actors and objects during the interaction process, as well as to specify the target object of interaction. To ensure a fair comparison, we re-implement and adapt several control-based baselines under our experimental settings. Across both datasets and multiple evaluation metrics, our method consistently outperforms all other baselines. This performance gain demonstrates the effective-

Method	Pub.	FVD↓	LPIPS↓	PSNR↑	SSIM↑	V2V-Sim↑	T2V-Sim↑
DynamiCrafter	ECCV24	554.48 / 860.53	0.516 / 0.375	13.48 / 14.21	0.553 / 0.571	0.473 / 0.867	0.146 / 0.215
DynamiCrafter-ft	ECCV24	168.73 / 197.82	0.206 / 0.166	20.49 / 19.80	0.721 / 0.775	0.814 / 0.957	0.199 / **0.223**
CosHand	ECCV24	162.87 / 174.84	0.209 / 0.123	20.67 / 21.81	0.725 / 0.809	0.837 / 0.969	0.191 / 0.220
InterDyn	CVPR25	172.42 / 207.80	0.207 / 0.145	20.71 / 21.16	0.730 / 0.802	0.794 / 0.955	0.172 / 0.219
Mask2IV (Ours)	This Work	**149.68 / 155.73**	**0.178 / 0.111**	**21.48 / 22.30**	**0.741 / 0.815**	**0.847 / 0.971**	**0.200** / 0.220

Table 1: Quantitative comparisons on HOI4D / BridgeData V2. DynamiCrafter and its fine-tuned variant perform image-to-video generation without explicit control, while the remaining methods incorporate control signals through the use of masks.

Figure 4: Qualitative comparison with CosHand and InterDyn. Generation artifacts are marked with yellow circles.

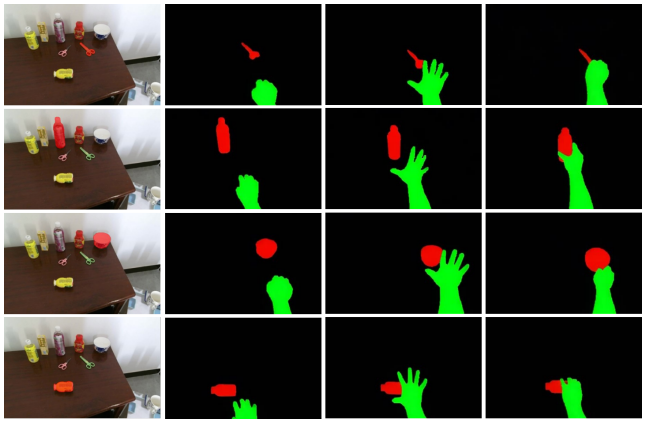

Figure 5: Qualitative analysis of interaction trajectory generation involving different objects. Target objects are highlighted in red in the initial image.

	FVD↓	LPIPS↓	PSNR↑	SSIM↑
ControlNet	157.38	0.182	21.49	0.747
MaskLatent	130.07	0.157	22.33	0.760
+ object mask	115.14	0.132	23.85	0.802
+ random d/e	108.80	0.124	24.16	0.802
+ contact loss	104.61	0.126	24.37	0.804

Table 2: Ablation results on the HOI4D dataset conducted with ground-truth mask trajectories. Random d/e denotes random dilation or erosion applied to the masks.

ness of our trajectory-conditioned design and joint modeling of actors and objects in capturing fine-grained, controllable interactions, while preserving video quality and temporal coherence.

Qualitative Analysis

In Fig. 4, we present visual comparisons between our method and two representative baselines. Our method consistently produces higher-quality interaction videos, characterized by coherent motion and clear contact regions. In contrast, the baseline methods often suffer from visual artifacts, such as inconsistent color patterns (*e.g.*, the white lid of a trashcan turning green) and incomplete object manipulation, where only parts of an object are affected (*e.g.*, the handle and the pot appear detached). Moreover, a notable advantage of our method lies in its flexible and versatile controllability. By simply modifying the input object mask, Mask2IV can synthesize diverse interaction sequences involving different objects. This capability is illustrated in Fig. 5, where different target objects are grasped under the same input image, showcasing strong generalization to object variation. We further analyze the controllability of our method under different conditioning prompts. As shown in Fig. 6, given the same input image, our framework allows users to adjust the text prompt or position mask to guide the generation process. This enables fine-grained control over the type of interaction, such as pushing or pulling a toy car, or opening or closing a laptop. It also supports spatial control over the object's placement, allowing the specification of both the location and orientation of the target object, which facilitates the generation of diverse and robust data for downstream robot learning tasks.

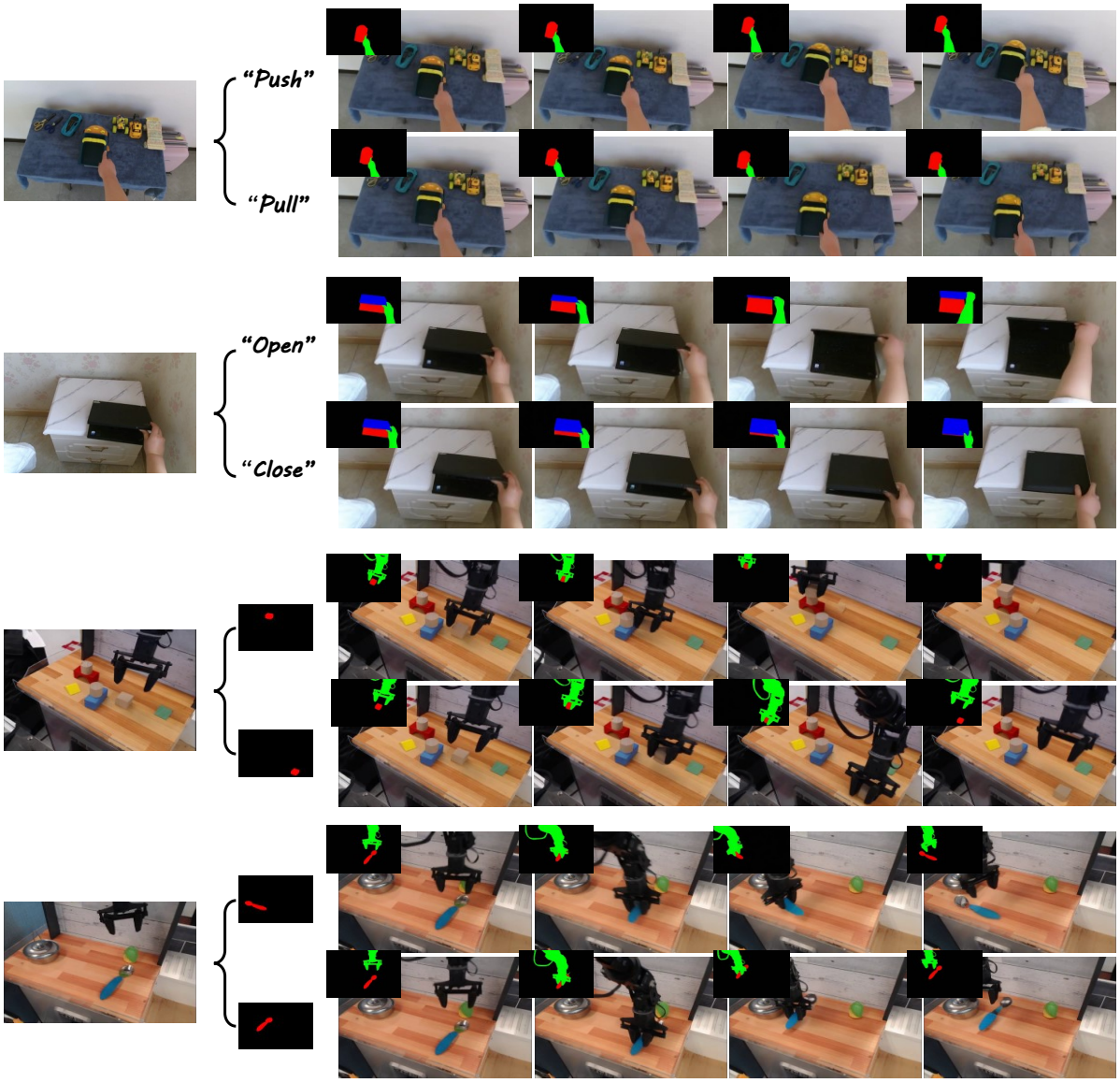

Figure 6: Qualitative analysis of different text prompts and position masks.

Ablation Study

An ablation study is performed in Tab. 2 to evaluate the contribution of each component. We first examine different control schemes and find that directly concatenating mask latent features with the noise input yields stronger performance than training an auxiliary ControlNet, leading to improved training stability and faster convergence in early stages. After integrating object masks into the trajectory, we observe consistent improvements across all metrics. These gains highlight the importance of modeling interactions through the motions of both actors and objects, rather than relying solely on actor representations. The application of random mask perturbations further boosts performance by enhancing the model's robustness to variations in mask shape during training. Lastly, the inclusion of the contact weighting loss leads to additional gains in video quality, particularly in accurately synthesizing interaction regions between the actor and the object.

Conclusion

In this paper, we introduce the task of interaction-centric video generation, focusing on synthesizing dynamic interactions between actors (humans or robots) and objects. We propose a novel two-stage pipeline that decouples interaction trajectories modeling from video synthesis, offering a more controllable and practical solution. To support training and evaluation in this domain, we further introduce two dedicated benchmarks covering both human-object interaction and robotic manipulation scenarios. Experimental results validate the effectiveness of our approach, highlighting its potential to advance controllable and realistic video generation in interactive settings.

Acknowledgments

This work is supported by a Start-up Grant from Nanyang Technological University and jointly funded by the Singapore Ministry of Education (MOE) under a Tier-1 research grant.

References

Akkerman, R.; Feng, H.; Black, M. J.; Tzionas, D.; and Abrevaya, V. F. 2025. InterDyn: Controllable Interactive Dynamics with Video Diffusion Models. *CVPR*.

Authors, G. 2024. Genesis: A Generative and Universal Physics Engine for Robotics and Beyond.

Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv e-prints*, arXiv–2305.

Dai, Z.; Zhang, Z.; Yao, Y.; Qiu, B.; Zhu, S.; Qin, L.; and Wang, W. 2023. AnimateAnything: Fine-Grained Open Domain Image Animation with Motion Guidance. *arXiv preprint arXiv:2311.12886*.

Fan, Y.; Yang, Q.; Wang, K.; Zhou, H.; Li, Y.; Feng, H.; Wu, Y.; and Wang, J. 2025. Re-HOLD: Video Hand Object Interaction Reenactment via adaptive Layout-instructed Diffusion Model. *arXiv preprint arXiv:2503.16942*.

Fang, G.; Yan, W.; Guo, Y.; Han, J.; Jiang, Z.; Xu, H.; Liao, S.; and Liang, X. 2024. Humanrefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, 201–217. Springer.

Furuta, H.; Zen, H.; Schuurmans, D.; Faust, A.; Matsuo, Y.; Liang, P.; and Yang, S. 2024. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*.

Geng, D.; Herrmann, C.; et al. 2024. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hoe, J. T.; Jiang, X.; Chan, C. S.; Tan, Y.-P.; and Hu, W. 2024. Interactdiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6180–6189.

Hu, H.; Wang, W.; Zhou, W.; and Li, H. 2022. Hand-object interaction image generation. *Advances in Neural Information Processing Systems*, 35: 23805–23817.

Jiang-Lin, J.-Y.; Huang, K.-Y.; Lo, L.; Huang, Y.-N.; Lin, T.; Wu, J.-C.; Shuai, H.-H.; and Cheng, W.-H. 2024. Record: Reasoning and correcting diffusion for hoi generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9465–9474.

Kareer, S.; Patel, D.; Punamiya, R.; Mathur, P.; Cheng, S.; Wang, C.; Hoffman, J.; and Xu, D. 2024. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*.

Kulal, S.; Brooks, T.; Aiken, A.; Wu, J.; Yang, J.; Lu, J.; Efros, A. A.; and Singh, K. K. 2023. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17089–17099.

Lai, B.; Dai, X.; Chen, L.; Pang, G.; Rehg, J. M.; and Liu, M. 2024. LEGO: L earning EGO centric Action Frame Generation via Visual Instruction Tuning. In *European Conference on Computer Vision*, 135–155. Springer.

Lepert, M.; Fang, J.; and Bohg, J. 2025. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*.

Li, G.; Jampani, V.; Sun, D.; and Sevilla-Lara, L. 2023a. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10922–10931.

Li, G.; Tsagkas, N.; Song, J.; Mon-Williams, R.; Vijayakumar, S.; Shao, K.; and Sevilla-Lara, L. 2025. Learning Precise Affordances from Egocentric Videos for Robotic Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, Y.; Cao, Z.; and Corso, J. J. 2024. HANDI: Hand-Centric Text-and-Image Conditioned Video Generation. *arXiv preprint arXiv:2412.04189*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22511–22521.

Liang, J.; Fan, Y.; Zhang, K.; Timofte, R.; Van Gool, L.; and Ranjan, R. 2024. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, 56–74. Springer.

Lin, K. Q.; et al. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35: 7575–7586.

Liu, S.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Liu, Y.; Liu, Y.; Jiang, C.; Lyu, K.; Wan, W.; Shen, H.; Liang, B.; Fu, Z.; Wang, H.; and Yi, L. 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21013–21022.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Luo, G. Y.; Luo, Z. H.; Gosselin, A.; Jolicoeur-Martineau, A.; and Pal, C. 2024. Ctrl-V: Higher Fidelity Video Generation with Bounding-Box Controlled Object Motion. *arXiv preprint arXiv:2406.05630*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Narasimhaswamy, S.; Bhattacharya, U.; Chen, X.; Dasgupta, I.; Mitra, S.; and Hoai, M. 2024. Handiffuser: Text-to-image generation with realistic hand appearances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2468–2479.

Park, J.; Kong, K.; and Kang, S.-J. 2024. AttentionHand: Text-Driven Controllable Hand Image Generation for 3D Hand Reconstruction in the Wild. In *European Conference on Computer Vision*, 329–345. Springer.

Qin, Z.; Zhang, Y.; Liu, Y.; and Campbell, D. 2024. HandCraft: Anatomically Correct Restoration of Malformed Hands in Diffusion Generated Images. *arXiv preprint arXiv:2411.04332*.

Radford, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ravi, N.; Gabeur, V.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, T.; Liu, S.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Shi, X.; et al. 2024. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Souček, T.; Damen, D.; Wray, M.; Laptev, I.; and Sivic, J. 2024. Genhowto: Learning to generate actions and state transformations from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6561–6571.

Sudhakar, S.; Liu, R.; Hoorick, B. V.; Vondrick, C.; and Zemel, R. 2024. Controlling the world by sleight of hand. In *European Conference on Computer Vision*, 414–430. Springer.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. FVD: A new metric for video generation. *International Conference on Learning Representations Workshop*.

Walke, H. R.; Black, K.; et al. 2023. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, 1723–1736. PMLR.

Wang, J.; Zhang, Y.; Zou, J.; Zeng, Y.; Wei, G.; Yuan, L.; and Li, H. 2024a. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*.

Wang, X.; Zhao, K.; Liu, F.; Wang, J.; Zhao, G.; Bao, X.; Zhu, Z.; Zhang, Y.; and Wang, X. 2024b. EgoVid-5M: A Large-Scale Video-Action Dataset for Egocentric Video Generation. *arXiv preprint arXiv:2411.08380*.

Wang, Y.; Xian, Z.; Chen, F.; Wang, T.-H.; Wang, Y.; Fragkiadaki, K.; Erickson, Z.; Held, D.; and Gan, C. 2024c. RoboGen: towards unleashing infinite data for automated robot learning via generative simulation. In *Proceedings of the 41st International Conference on Machine Learning*.

Wang, Y.; et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024d. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2024. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, 399–417. Springer.

Xu, Z.; Chen, Q.; Peng, Y.; and Liu, Y. 2024a. Semantic-aware human object interaction image generation. In *Forty-first International Conference on Machine Learning*.

Xu, Z.; Huang, Z.; Cao, J.; Zhang, Y.; Cun, X.; Shuai, Q.; Wang, Y.; Bao, L.; Li, J.; and Tang, F. 2024b. AnchorCrafter: Animate CyberAnchors Saling Your Products via Human-Object Interacting Video Generation. *arXiv preprint arXiv:2411.17383*.

Xue, Z. S.; Luo, R.; Chen, C.; and Grauman, K. 2024. Hoiswap: Swapping objects in videos with hand-object interaction awareness. *Advances in Neural Information Processing Systems*, 37: 77132–77164.

Yang, C.; Kang, C.; Kong, K.; Oh, H.; and Kang, S.-J. 2024. Person in Place: Generating Associative Skeleton-Guidance Maps for Human-Object Interaction Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8164–8175.

Yariv, G.; Kirstain, Y.; Zohar, A.; Sheynin, S.; Taigman, Y.; Adi, Y.; Benaim, S.; and Polyak, A. 2025. Through-The-Mask: Mask-based Motion Trajectories for Image-to-Video Generation. *arXiv preprint arXiv:2501.03059*.

Ye, Y.; et al. 2023. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22479–22489.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhen, H.; Sun, Q.; Zhang, H.; Li, J.; Zhou, S.; Du, Y.; and Gan, C. 2025. TesserAct: learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*.