This CVPR 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.

PAM: A Pose–Appearance–Motion Engine for Sim-to-Real HOI Video Generation

Mingju Gao*[1,2], Kaisen Yang*[2], Huan-ang Gao[2], Bohan Li[4,5], Ao Ding[2],
Wenyi Li[2], Yangcheng Yu[2], Jinkun Liu[2], Shaocong Xu[3],
Yike Niu[2], Haohan Chi[2], Hao Chen[6], Hao Tang[1], Yu Zhang[†4], Li Yi[2], Hao Zhao[†2,3]
[1]Peking University [2]Tsinghua University [3]BAAI [4]SJTU
[5]Eastern Institute of Technology [6]University of Cambridge
Project Page: https://gasaiyu.github.io/PAM.github.io/

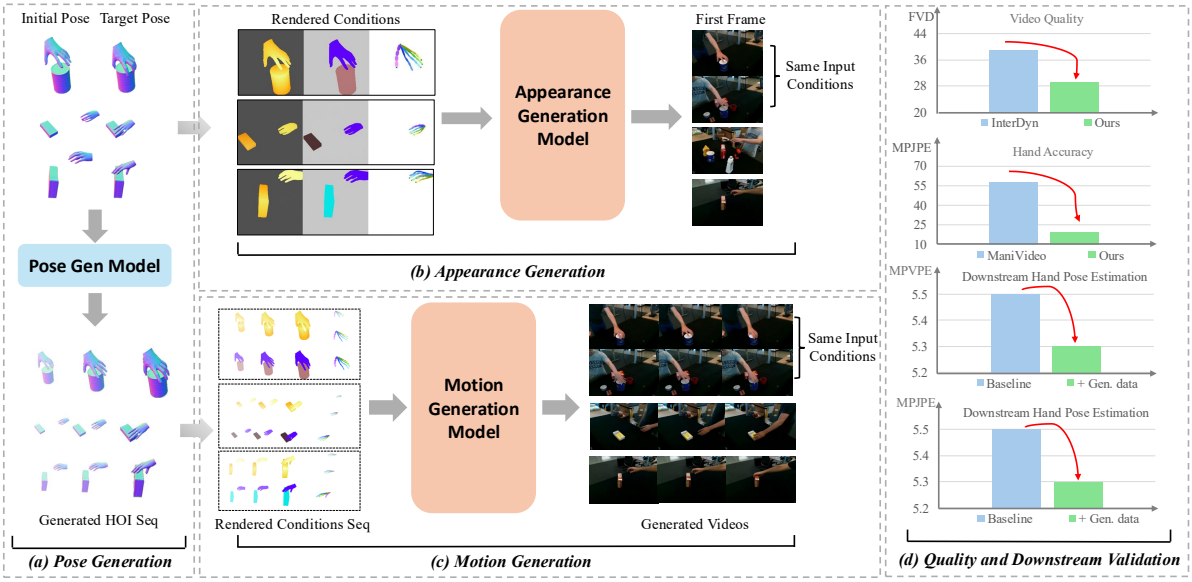

Figure 1. **(a) Pose Generation.** Given an initial and target hand–object pose, our pose generation module synthesizes a full MANO-based HOI trajectory. **(b) Appearance Generation.** Conditioned on rendered hand–object geometry, the appearance model produces diverse and realistic reference frame. **(c) Motion Generation.** Using the same rendered conditions, the motion model generates the full HOI video with coherent dynamics and hand–object interactions. **(d) Quality and Downstream Validation.** Our unified engine substantially improves video fidelity and geometric accuracy over prior methods.

Abstract

Hand–object interaction (HOI) reconstruction and synthesis are becoming central to embodied AI and AR/VR. Yet, despite rapid progress, existing HOI generation research remains fragmented across three disjoint tracks: (1) pose-only synthesis that predicts MANO trajectories without producing pixels; (2) single-image HOI generation that hallucinates appearance from masks or 2D cues but lacks dynamics; and (3) video generation methods that require both the entire pose sequence and the ground-truth first frame as inputs, preventing true sim-to-real deployment. Inspired by the philosophy of [33], we think that HOI generation requires a unified engine that brings together pose, appearance, and motion within one coherent framework. Thus we introduce PAM: a Pose–Appearance–Motion Engine for controllable HOI video generation. The performance of our engine is validated by: (1) On DexYCB, we obtain an FVD of 29.13 (vs. 38.83 for InterDyn), and MPJPE of 19.37 mm (vs. 30.05 mm for CosHand), while generating higher-resolution 480×720 videos compared to 256×256/256×384 baselines. (2) On OAKINK2, our full multi-condition model improves FVD from 68.76 → 46.31. (3) An ablation over input conditions on DexYCB shows that combining depth, segmentation, and keypoints consistently yields the best results. (4) For a downstream hand pose estimation task using SimpleHand, augmenting training with 3,400 synthetic videos (207k frames) allows a model trained on only 50% of the real data plus our synthetic data to match the 100% real baseline.

[0]*Equal Contribution. †Corresponding Author.

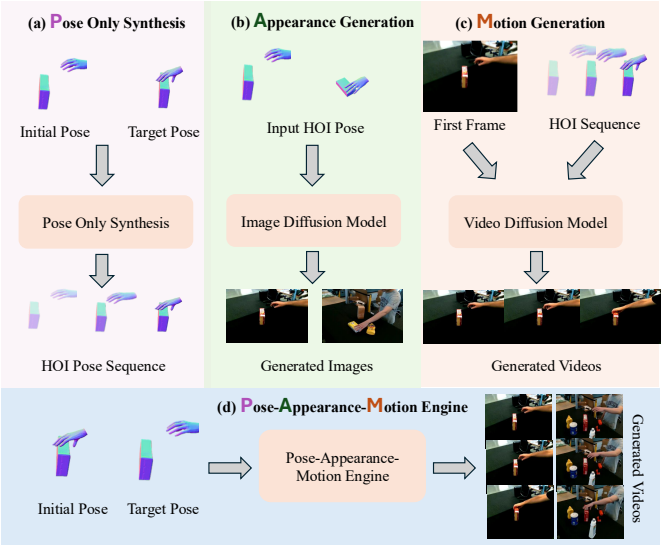

Figure 2. **Overview of Four Approaches to HOI Synthesis. (a) Pose-Only Synthesis [78]:** This method predicts the MANO trajectories without generating pixel data; **(b) Apperance Generation [82]:** This approach generates appearance based on masks or 2D cues but lacks dynamic motion; **(c) Motion Generation [3, 48]:** These methods require both the full pose sequence and the ground-truth first frame as inputs, limiting their application for true sim-to-real transfer. **(d) Our Pipeline PAM:** In this approach, video generation does not rely on the first frame or the whole HOI pose sequence, allowing for the transfer of HOI pose sequences from the simulator to real-world videos.

1. Introduction

The ability to manipulate objects with our hands represents a fundamental human skill, and the computational understanding of this capability—referred to as Hand-Object Interaction (HOI) understanding—has become increasingly significant in the fields of computer vision and embodied AI. The field has seen a shift towards **data-driven** paradigms, where large-scale HOI datasets are instrumental for accurate hand pose estimation [32, 55, 64, 74, 75, 79, 92], enabling realistic human-to-robot motion transfer [38, 40, 45] and 3D modeling [9–11, 18, 19, 44, 52, 57, 65, 75, 84, 86]. The critical challenge, however, lies in the data itself. Despite considerable investments in collecting real-world HOI sequences with detailed annotations [20, 46, 69], the reliance on costly and labor-intensive manual labeling poses a fundamental limitation to scalability.

The advent of deep learning and diffusion models has opened up promising avenues for scalable generation of HOI videos. However, as illustrated in Figure 2, state-of-the-art methods [3, 48, 60, 78, 82, 89] still face significant challenges: (1) Pose-only synthesis approaches are limited in generating realistic appearances, as they predict MANO trajectories without producing pixel-level details, which diminishes their practical utility. (2) Single-image HOI generation (appearance generation) methods, which rely on hallucinating appearance from masks or 2D cues, fail to capture temporal dynamics, resulting in incoherent motion when applied to real-world scenarios. (3) Video generation (motion generation) methods that require both the complete pose sequence and the ground-truth first frame are unsuitable for sim-to-real deployment, as ground-truth first frame is unavailable from the simulator. These limitations emphasize the need for innovative approaches that can generate high-quality, temporally consistent HOI videos without depending on fixed pose sequences or ground-truth first frame inputs, thus providing a more flexible and robust solution for sim-to-real transfer.

In this paper, we introduce PAM: a pioneering Pose–Appearance–Motion engine for sim-to-real HOI video generation that requires *only* the initial and target poses, along with object geometry, as input. By integrating a motion-appearance diffusion process, our method bypasses the need for a conditioned first frame, thereby maximizing both motion and appearance diversity—a capability unattainable by prior work. To ensure high realism, we incorporate multiple conditions that effectively preserve fine hand details.

Our framework consists of three core stages. The **pose sequence generation** stage produces a plausible hand motion trajectory using a pre-trained model. Next, the **appearance generation** stage synthesizes a realistic initial frame using the controllable image diffusion model Flux [5]. This model is conditioned on a fusion of depth maps, semantic masks, and hand keypoint maps, ensuring geometric accuracy, semantic coherence, and the preservation of fine-grained hand details. The **motion generation** stage first renders the hand motion trajectory into depth, semantic, and hand keypoint sequences, and then generates videos using a controllable video diffusion model based on CogVideoX [70]. Importantly, the video model is conditioned on the same multi-modal inputs as the appearance generation stage to ensure consistency with the generated HOI pose sequence.

We evaluate our method on the DexYCB [7] and OAKINK2 [76] benchmarks, where it comprehensively surpasses existing approaches in video generation quality, motion plausibility, and hand pose fidelity. More importantly, as evidenced in Figure 1, the synthesized videos from our method provide substantial value as synthetic data. When used for training, they lead to meaningful gains in the performance of a downstream hand pose estimation model, demonstrating their effectiveness as a data augmentation tool.

We summarize our contributions as follows:

- **Minimal-Conditioning Generation:** We pioneer a pose–appearance–motion engine for sim-to-real HOI video generation that requires only sparse pose keyframes and object geometry as input, overcoming the first-frame bottleneck of prior methods.
- **Decoupled Generation Architecture:** We design a novel pipeline that decouples pose, appearance and motion syn-

thesis, leveraging multi-modal conditions to achieve superior realism, controllability and diversity.

- **State-of-the-Art Performance and Utility:** Our method achieves superior results on established benchmarks and proves its practical value by enabling significant gains in downstream task performance through effective data augmentation.

2. Related Works

2.1. Hand-Object Motion Synthesis

Synthesizing high-fidelity hand-object motion is a fundamental challenge in computer animation and robotic grasping [1, 13, 21, 41, 87]. Prevailing data-driven approaches rely on supervised learning from large-scale, well-annotated datasets [12, 15, 30, 31, 35, 40, 43, 49, 53, 85, 90–92]. However, the scalability of these methods is constrained by their dependence on costly and difficult-to-acquire data [7, 17, 20, 26, 46, 47, 69, 76]. To circumvent this limitation, reinforcement learning (RL) has emerged as a promising alternative. Methods like [13, 68] generate reference grasps before synthesizing motions, while GraspXL [78] learns a generalizable grasping policy directly in simulation, eliminating the need for predefined references. These RL-based techniques produce high-quality interaction data, forming a robust foundation for sim-to-real transfer. However, these methods lack appearance modeling.

2.2. Controllable Video Generation

Recent breakthroughs in video generation foundation models [2, 6, 36, 62, 70] have intensified interest in controllable generation that precisely aligns with user intent. While text-to-video and image-to-video models [2, 24, 39, 54, 58, 62, 67, 70, 80] have demonstrated impressive capabilities, they often lack the granularity for specialized tasks. This has spurred research into integrating more precise control signals, such as semantic maps, depth, and camera motion. ControlNet [81] and its variants [22, 25, 42, 77] enable conditioning on dense inputs, while works like VideoComposer [66] fuse multiple conditions for enhanced control. Camera motion has been explicitly modeled by embedding parameters into diffusion models [4, 27]. However, generating videos of hand-object interactions (HOI) presents a unique challenge due to the high degrees of freedom in hand motion. This demands even more enriched and specialized control mechanisms—combining semantic, geometric, and precise pose cues—to achieve the necessary fidelity and accuracy.

2.3. Hand-Object Interaction Image & Video Generation

Generating Hand-Object Interaction (HOI) content is vital for understanding human activities. Prior work on *HOI image generation* [8, 28, 37, 51, 63, 73, 82] typically conditions on 2D signals like segmentation masks and keypoints. However, these static methods cannot capture the dynamic nature of interactions. Recently, several studies [3, 14, 16, 48, 60, 72, 88] have explored *HOI video generation*. InterDyn [3] conditions on hand mask sequences via ControlNet [81], but under-utilizes the rich conditions available from simulators. ManiVideo [48] introduces an occlusion-aware representation but requires human appearance data, which is not available from simulators like GraspXL [78]. More critically, these methods primarily focus on generation quality and have not thoroughly investigated the *downstream utility* of their synthesized data, which is essential for validating practical impact beyond perceptual metrics.

3. The Proposed Method

3.1. Overview

Figure 3 illustrates our method. Conditioned on an initial MANO [56] hand pose $\mathbf{h}_0 \in \mathbb{R}^{51\times 3}$, an object mesh $\mathbf{m}$ without appearance, an initial 6-DoF object pose $\mathbf{o}_0 \in \mathbb{R}^6$, and a target hand pose $\mathbf{h}_T \in \mathbb{R}^{51\times 3}$, our generative model

$$f_{\boldsymbol{\theta}} : (\mathbf{h}_0, \mathbf{m}, \mathbf{o}_0, \mathbf{h}_T) \;\rightarrow\; \{I_t\}_{t=0}^{T} \tag{1}$$

produces a photo-realistic video that (i) begins with $\mathbf{h}_0$, (ii) ends with $\mathbf{h}_T$, and (iii) depicts a temporally-coherent grasp-to-place motion. All hand poses are parameterised by global translation + rotation plus joint angles; frames I_t are RGB images.

Jointly modeling pose, appearance and motion is inherently challenging due to the high-dimensional spatio-temporal manifold [23]. To address this, we decompose the generation process into three distinct stages:

1. **Pose Generation** — A pretrained pose generation model synthesizes aligned hand-object pose sequences, $\{\mathbf{h}_t, \mathbf{o}_t\}_{t=0}^{T}$, from the initial and target hand-object poses and object mesh, $(\mathbf{h}_0, \mathbf{o}_0, \mathbf{m})$ (Sec. 3.2).
2. **Appearance Generation** — A controllable image diffusion model generates the first frame, I_0, conditioned on the initial HOI pose and object mesh (Sec. 3.3).
3. **Motion Generation** — The synthetic HOI pose sequences and the first frame are injected into a video diffusion model, which animates I_0 into a photo-realistic video clip $\{I_t\}_{t=0}^{T}$ (Sec. 3.4).

3.2. Stage I: Pose Generation Stage

In the first stage, we aim to generate an interpolated hand-object interaction (HOI) pose sequence, transitioning from the initial to the target pose while incorporating the object mesh. As shown in Figure 3, we employ GraspXL [78], a pretrained model for hand-object interaction tasks. GraspXL takes as input the initial and target MANO hand pose $\mathbf{h}_0$, the 6-DoF object pose $\mathbf{o}_0$, and the object mesh $\mathbf{m}$, and produces

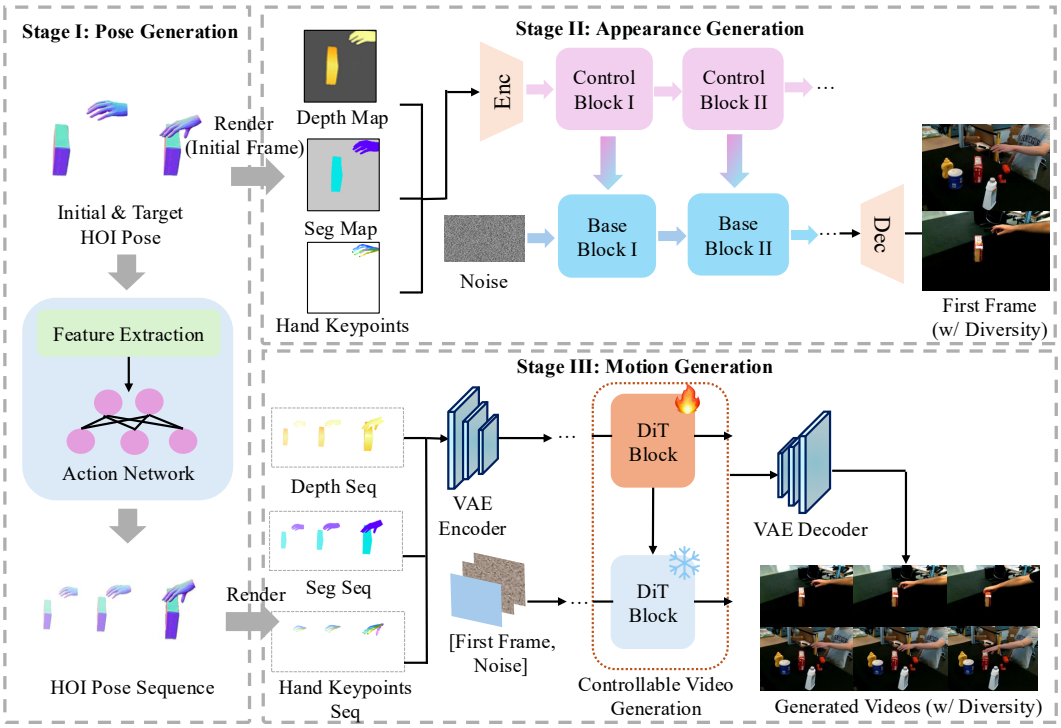

Figure 3. **Overview of our three-stage generation pipeline.** (1) **Pose Generation:** A pretrained pose generation model generates the intermediate hand-object interaction (HOI) poses based on the initial and target poses, along with the object mesh. (2) **Appearance Generation:** A controllable image diffusion model synthesizes the first frame of the video, conditioned on multi-modal inputs (depth maps, semantic masks, and keypoint annotations). (3) **Motion Generation:** The generated HOI sequence and the first frame are rendered into a full video sequence by a video diffusion model, conditioned on the same multi-modal inputs used in the appearance generation stage.

temporally coherent trajectories $\{\mathbf{h}_t, \mathbf{o}t\}_{t=0}^{T}$ for both hand and object poses. This ensures smooth interpolation while maintaining the physical consistency of the hand-object interaction, providing the foundation for subsequent video generation stages.

3.3. Stage II: Appearance Generation Stage

Bridge Conditions for Sim-to-Real HOI Video Synthesis: The primary objective of this work is to enhance the visual quality of simulated videos while preserving other conditions, thereby bridging the gap between simulation and real-world scenarios. By incorporating both geometric information (e.g., depth maps) and semantic data (e.g., segmentation masks) from the simulator, we seek to accurately reconstruct the visual representation of scenes and objects, while ensuring consistency across all other conditions. However, relying solely on these two data types proves insufficient for accurately generating Hand-Object Interactions (HOI) images or videos. This limitation stems from the complexity and high degree of freedom inherent in hand movements, which cannot be fully captured by geometric and semantic data alone. Specifically, these conditions fail to account for critical details, such as the number of fingers and their individual poses. To address this challenge, we introduce an additional condition—hand keypoint sequences, as proposed by [82]—to enable more precise and accurate hand pose generation. This approach facilitates the generation of realistic hand poses, thereby enhancing the overall realism of the interaction. In section 4.1 and 4.2, we explore the influence of every condition.

We fine-tune Flux [5] with a ControlNet [81] fork that accepts depth D_0, segmentation S_0 and hand-keypoint image K_0 ($H{\times}W{\times}3$ each). All cues are VAE-encoded to $\frac{H}{8}{\times}\frac{W}{8}{\times}16$ latents, concatenated channel-wise and Injected into two layers of DiT [50] blocks, with weights initialized from the first two layers of original Flux.:

$$f_l = f_l + \mathcal{Z}(f_l'), \tag{2}$$

where f_l is the output of the l-th layer of the original DiT [50] blocks, and f_l' is the output of the l-th layer of the duplicated DiT blocks whose input is the concatenated conditions. Here, $l \in \{0, 1\}$, and $\mathcal{Z}$ represent the zero-convolution layer, which is a 1×1 convolution with all parameters initialized to zero. During training, only the parameters of ControlNet are updated.

3.4. Stage III: Motion Generation Stage

To generate the target video sequence, we combine the generated HOI pose sequence with a controllable video diffusion model. As shown in Figure 3, depth maps, instance-level

segmentation masks, and 2-D hand keypoint images are rasterized at each frame from Stage I. These conditions are then encoded into a latent tensor of shape $\mathbb{R}^{\frac{T+1}{4} \times \frac{H}{8} \times \frac{W}{8} \times 16}$ using a pretrained video VAE. Next, the latent representations are concentrated along channel dimensions and injected into CogVideo-X through 12 duplicate DiT blocks, as described in Eq. 2. To prevent over-reliance on any single modality during training, each cue is randomly masked with a probability of 0.2.

4. Experiment

4.1. Experiment Settings

Datasets and Data Processing. We evaluate our method on two standard benchmarks for HOI video generation: DexYCB [7] and OAKINK2 [76]. For DexYCB, we adopt the s0-split, comprising 6,400 training and 1,600 validation videos. Due to the scale of OAKINK2, we use a curated subset of 8,000 video clips (each 49 frames long), split into 6,400 for training and 1,600 for validation. The conditions for our model—depth maps, semantic masks, and hand keypoints—are derived as follows: depth maps are estimated using DepthCrafter [29], while semantic and keypoint information are obtained directly from the dataset annotations.

Evaluation Metrics. We employ a comprehensive set of metrics to evaluate our method from four perspectives:

- **Image Quality:** We assess perceptual quality using Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) [83] and Peak Signal-to-Noise Ratio (PSNR).
- **Spatio-temporal Coherence:** We adopt Fréchet Video Distance (FVD) [61] to evaluate overall video realism, using the implementation from [59].
- **Motion Fidelity:** We use the Motion Fidelity (MF) metric [71] to quantify dynamic accuracy. For each video, we sample 100 foreground points (on hands/objects), track them using CoTracker3 [34], and compare the trajectories between generated and ground-truth videos. For a ground-truth tracklet $\mathcal{T} = \{\tau_1, \dots, \tau_T\}$ and a generated tracklet $\tilde{\mathcal{T}} = \{\tilde{\tau}_1, \dots, \tilde{\tau}_T\}$ where $\tau_t \in \mathbb{R}^2$, MF is defined as:

$$\text{MF} = \frac{1}{|\tilde{\mathcal{T}}|} \sum_{\tilde{\tau} \in \tilde{\mathcal{T}}} \max_{\tau \in \mathcal{T}} \mathbf{corr}(\tau, \tilde{\tau}) + \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \max_{\tilde{\tau} \in \tilde{\mathcal{T}}} \mathbf{corr}(\tau, \tilde{\tau}). \tag{3}$$

  The correlation between two tracks is computed as:

$$\mathbf{corr}(\tau, \tilde{\tau}) = \frac{1}{F} \sum_{k=1}^{F} \frac{\mathbf{v}_k \cdot \tilde{\mathbf{v}}_k}{\|\mathbf{v}_k\| \|\tilde{\mathbf{v}}_k\|}, \tag{4}$$

  where $\mathbf{v}_k = (v_k^x, v_k^y)$ and $\tilde{\mathbf{v}}_k = (\tilde{v}_k^x, \tilde{v}_k^y)$ are the displacement vectors at the k-th frame for tracks τ and $\tilde{\tau}$, respectively.
- **Hand Pose Accuracy:** We report Mean Per-Joint Position Error (MPJPE) in millimeters [17], measuring the average Euclidean distance between the 21 predicted and ground-truth hand joints after root alignment. Lower MPJPE indicates better pose estimation accuracy. We utilize Hamer [49] to estimate the hand joints in the generated videos

4.2. Main Results

Baselines. We compare our method against state-of-the-art HOI video generation approaches: ManiVideo [48], InterDyn [3], and CosHand [60] on the DexYCB dataset [7]. For ManiVideo and InterDyn, we report results directly from their original publications (omitting metrics for which results were unavailable due to these methods not being open-source). For CosHand, we use the official implementation and fine-tune it on the DexYCB s0-split training set for a fair comparison. We also evaluate on OAKINK2, comparing against a similarly fine-tuned CosHand model. All baselines are evaluated in an image-to-video setting where the ground-truth first frame is provided, as this is required by these methods.

Quantitative Comparisons. Our quantitative evaluation (Tables 1, 2) demonstrates that our method achieves state-of-the-art performance across most metrics. We attribute this superiority to our multi-conditioning strategy, which provides the diffusion model with rich geometric and semantic cues (depth, masks, keypoints) to jointly optimize for visual realism and pose accuracy. In contrast, baseline methods exhibit limitations: InterDyn, ManiVideo, and CosHand rely on more limited conditioning signals or are built upon foundation models that struggle to capture the intricacies of hand-object interactions, leading to suboptimal performance.

Qualitative Comparisons. As shown in Figure 4, our method generates visually superior results compared to CosHand, even when CosHand is fine-tuned on the same training data. We identify two primary limitations in CosHand: (1) its reliance on hand masks as the sole conditioning signal provides insufficient geometric guidance for reconstructing precise hand poses, and (2) its lack of explicit temporal modeling mechanisms leads to inconsistent frame-to-frame outputs. In contrast, our approach addresses these issues by leveraging a video diffusion foundation model equipped with temporal attention to enforce coherence across frames. Furthermore, the use of hand keypoint maps as a conditioning signal explicitly preserves the structural details of hand configurations, resulting in more accurate and smooth video sequences.

4.3. Ablation Studies on Input Conditions

We conduct an ablation study on the DexYCB dataset to evaluate the contribution of different input conditions. The results (Tables 1, 2, and 3) yield three key observations:

- The performance improves with an increasing number of conditions, validating the effectiveness of our multi-

Method	Venue	FVD (↓)	MF (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	MPJPE (↓)	Resolution
CosHand [60]	ECCV'24	58.51	0.591	0.139	0.767	23.20	30.05	256 x 256
InterDyn [3]	CVPR'25	38.83	0.680	0.119	0.848	24.86	-	256 x 384
ManiVideo [48]	CVPR'25	-	-	0.079	0.913	30.10	57.30	-
Ours w/ all	-	**29.13**	**0.712**	**0.069**	**0.914**	**30.17**	**19.37**	480 x 720

Table 1. **Quantitative comparison on DexYCB dataset.** Our method is evaluated against CosHand, InterDyn, and ManiVideo. Results for InterDyn and ManiVideo are taken from their original papers. For fair comparison, CosHand was fine-tuned on the s0-split training set identical to ours. Our approach achieves state-of-the-art performance across all metrics (FVD, LPIPS, MF, MPJPE) while generating high-resolution 480x720 videos.

Method	FVD (↓)	MF (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	MPJPE (↓)
CosHand [60]	68.76	0.651	0.156	0.765	23.84	14.49
Ours w/ seg	48.97	0.708	0.084	0.831	25.76	9.61
Ours w/ depth	50.85	0.702	0.086	0.845	26.98	10.07
Ours w/ hand	52.41	0.671	0.113	0.838	25.66	8.01
Ours w/ all	**46.31**	**0.777**	**0.081**	**0.851**	**28.36**	**7.01**

Table 2. **Quantitative results on the OAKINK2 dataset.** Comparison of our method with CosHand. For a fair evaluation, both models are trained on the same dataset. Our approach achieves state-of-the-art performance, outperforming CosHand across all evaluated metrics.

Conditions	FVD (↓)	MF (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	MPJPE (↓)
Seg	33.23	0.695	0.077	0.900	29.27	21.14
Depth	30.00	0.703	0.070	0.906	29.15	23.16
Hand	33.41	**0.713**	0.086	0.901	29.07	20.70
Depth, Hand	29.62	0.711	0.071	0.899	29.95	20.46
Seg, Hand	29.53	0.711	0.073	0.902	29.57	19.92
Depth, Seg	29.32	0.712	0.071	0.906	**30.60**	22.51
Depth, Hand, Seg	**29.13**	0.712	**0.069**	**0.914**	30.17	**19.37**

Table 3. Ablation study on input conditions on DexYCB dataset.

condition design.

- Even when using the same segmentation mask condition as CosHand and InterDyn, our method achieves superior results, demonstrating the advantage of our pipeline.
- While using only hand keypoints yields low MPJPE (due to explicit pose supervision), it underperforms on other metrics due to the lack of broader geometric and semantic context. This highlights the necessity of combining detailed local cues (keypoints) with global scene understanding (depth, semantics) for optimal performance.

Visual results in Figure 5 further support these findings: using all conditions produces accurate poses; semantic masks or depth maps alone lead to pose inaccuracies; and keypoints alone degrade appearance quality.

4.4. Diverse Sim-to-Real Transfer

We conduct a sim-to-real transfer experiment to validate the effectiveness of our full pipeline. For this, we use GraspXL [78] as the hand motion generator, owing to its superior performance and strong generalization capabilities. Using objects from the DexYCB dataset, we randomly initialize both the hand and object poses, along with a target hand pose. GraspXL then generates the intermediate motion sequence, which is used to render the necessary conditions—depth, semantic masks, and keypoints—for our video generation model. As demonstrated in Figures 1 and 6, our method successfully synthesizes **diverse, realistic videos** with varying subjects and backgrounds, using minimal input that includes only the initial and target poses, along with the object geometry. This capability arises from our decoupled generation architecture, which effectively combines the motion prior from GraspXL with the appearance modeling of our diffusion model. The utility of these synthesized videos for downstream tasks is explored in Section 4.5.

4.5. Downstream Task Validation

To evaluate the utility of our generated videos, we employ them for data augmentation in a hand pose estimation task. We use SimpleHand [92] as the pose estimation model, which regresses MANO parameters [56] from a single image. Our pipeline, trained on DexYCB, generates 3,400 video sequences (207,400 frames) for augmentation. We combine this synthetic data with varying subsets (25%, 50%, 75%, 100%) of the original DexYCB s0-split training set (406,888 frames). All models are evaluated on the DexYCB validation set using four metrics: Procrustes-Aligned Mean Per-Joint Position Error (PA-MPJPE), Procrustes-Aligned Mean Per-Vertex Position Error (PA-MPVPE), and F-Score.

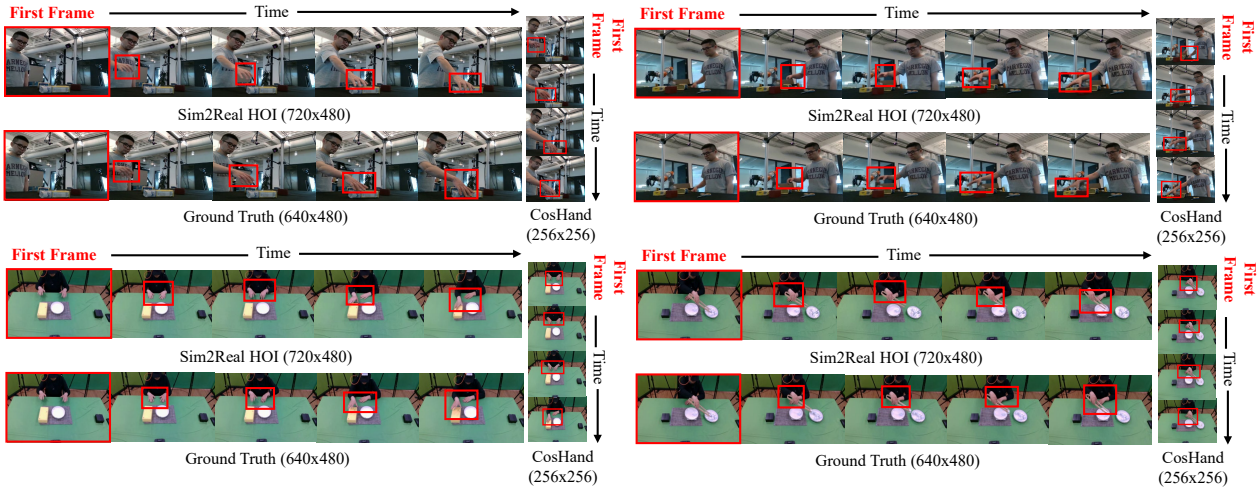

Figure 4. **Qualitative comparison against CosHand.** Example results on DexYCB and OAKINK2 highlight the strengths of our method in two key areas: (1) higher visual fidelity in both foreground and background generation, and (2) improved geometric accuracy of the synthesized hand poses.

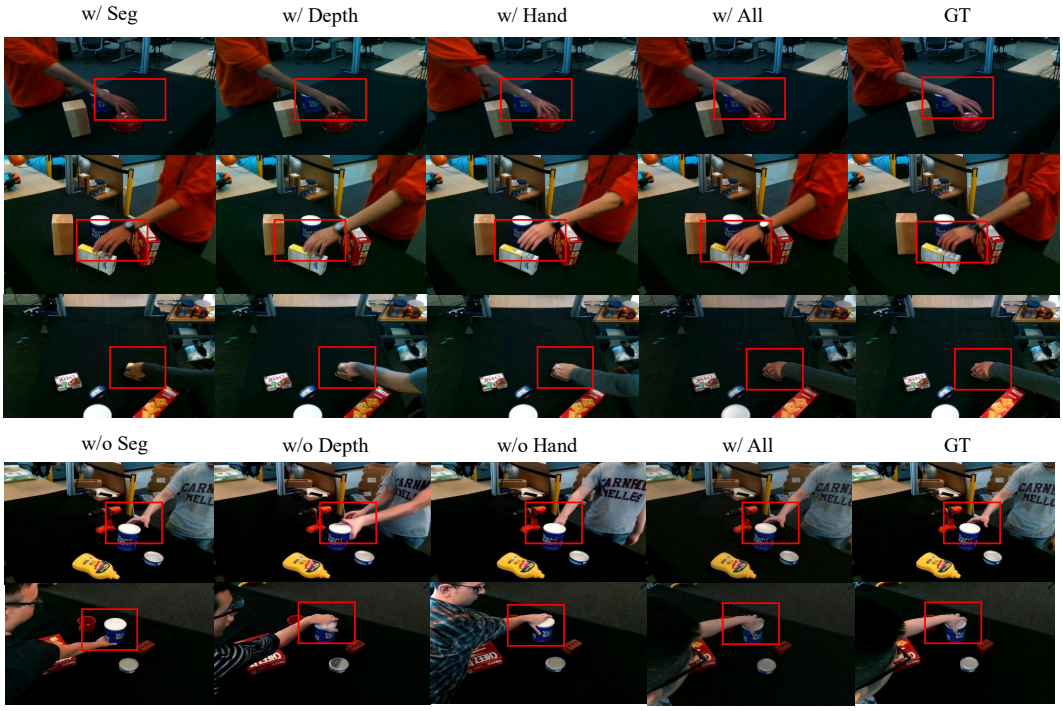

Figure 5. **Ablation study on input conditions on DexYCB dataset.**

PA-MPJPE/PA-MPVPE measure the average Euclidean distance (in mm) after Procrustes alignment between the predicted and ground-truth joints/vertices, respectively. The PA-MPJPE metric used here differs from the one in Table 1. The MPJPE in this context measures data efficiency for downstream tasks, whereas the MPJPE in Table 1 refers to the hand accuracy of the generated videos.

The quantitative results (Table 4) demonstrate that incorporating our generated data consistently improves hand pose estimation accuracy across all metrics. Figure 7 reveals two key trends: (1) model performance improves monotonically with the amount of real data, and (2) most notably, using only 50% of the real data augmented with our synthetic samples achieves competitive performance with the 100% real data baseline. This indicates that our synthetic data can effectively compensate for reduced real data volume. Furthermore, the superior performance achieved using videos generated with multiple conditions validates the importance of our multi-conditioning approach for producing diverse and useful training data.

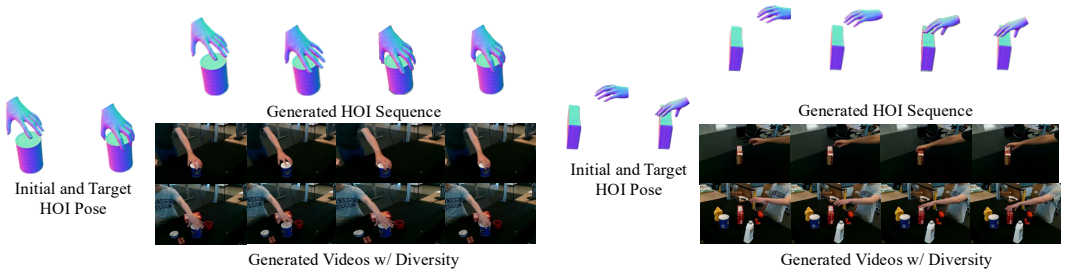

Figure 6. **Sim-to-real transfer results.** Our pipeline can generate realistic videos given initial and target states with diversity.

Setting	PA-MPJPE (↓)	PA-MPVPE (↓)	F-Score@05 (↑)	F-Score@15(↑)
All real data	5.5	5.5	0.7953	0.9899
All gen. data	8.2	8.1	0.6274	0.9626
All gen. + 25% real data	6.1	6.0	0.7512	0.9851
All gen. + 50% real data	5.5	5.5	0.8001	0.9879
All gen. + 75% real data	5.4	**5.3**	0.7984	0.9899
All gen. + 100% real data	**5.3**	**5.3**	**0.8025**	**0.9904**

Table 4. Downstream task evaluation on SimpleHand [92].

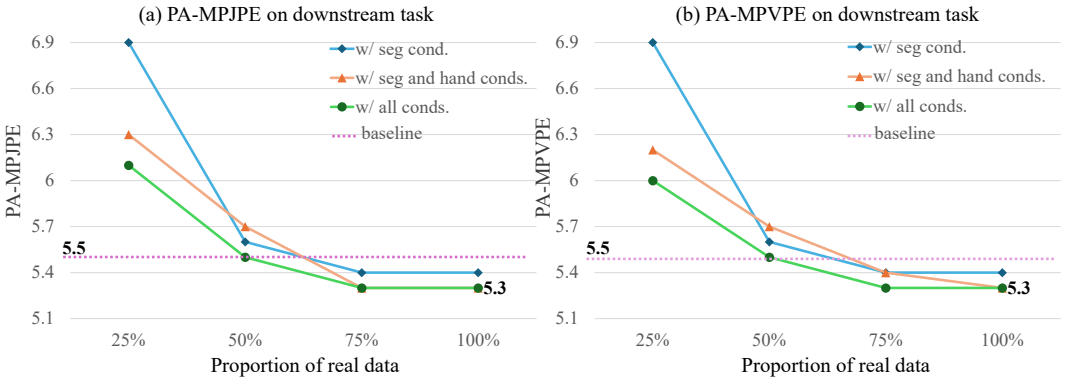

Figure 7. Data augmentation analysis with varying ratios of real data. We augment different portions of the DexYCB training set (25%, 50%, 75%, 100%) with our generated synthetic data. The **baseline** (dashed line) indicates performance when training solely on 100% of the real DexYCB data without synthetic augmentation.

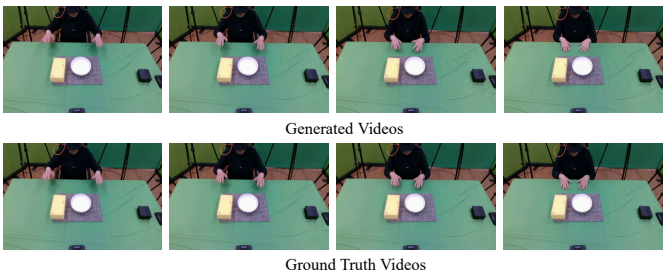

Figure 8. Zero-shot result on OAKINK2 dataset for i2v task. We use the weight trained on DexYCB dataset.

4.6. Zero-Shot Results

To evaluate the generalizability of our approach, we test our video diffusion model trained on the DexYCB dataset (single-hand interactions) directly on the OAKINK2 dataset (bimanual interactions) in a zero-shot image-to-video setting. As shown in Figure 8, our method generates plausible videos that maintain reasonable alignment with ground-truth hand poses and visual details, despite the significant domain shift. This cross-dataset generalization capability can be attributed to our use of pretrained video diffusion model weights as a strong foundation, combined with the ControlNet mechanism [81], which helps preserve the model's original generation quality while adapting to new conditioning signals.

5. Conclusion

This paper proposed a framework that addresses the challenge of generating realistic HOI videos from minimal pose inputs. Our decoupled, multi-condition architecture produces superior results in both perceptual quality and geometric accuracy, and demonstrates practical utility through enhanced downstream task performance. While our method shows strong generalization, future work could explore extending it to more complex object interactions or unifying the motion and appearance stages into an end-to-end model. We believe our contributions provide a solid foundation for future research in generative models for embodied AI.

Acknowledgments

This study is partially supported by Beijing Natural Science Foundation, QY25047.

References

[1] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. *arXiv preprint arXiv:2312.02975*, 2023. 3

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3

[3] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12467–12479, 2025. 2, 3, 5, 6

[4] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 3

[5] black-forest labs. https://github.com/black-forest-labs/flux, 2024. 2, 4

[6] 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. 3

[7] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021. 2, 3, 5

[8] Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17448–17460, 2025. 3

[9] Xiaoxue Chen, Tianyu Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Cerberus transformer: Joint semantic, affordance and attribute parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19649–19658, 2022. 2

[10] Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. Handos: 3d hand reconstruction in one stage. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17304–17314, 2025.

[11] Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. Hort: Monocular hand-held objects reconstruction with transformers. *arXiv preprint arXiv:2503.21313*, 2025. 2

[12] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 3

[13] Sammy Christen, Lan Feng, Wei Yang, Yu-Wei Chao, Otmar Hilliges, and Jie Song. Synh2r: Synthesizing hand-object motions for learning human-to-robot handovers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3168–3175. IEEE, 2024. 3

[14] Lingwei Dang, Ruizhi Shao, Hongwen Zhang, Wei Min, Yebin Liu, and Qingyao Wu. Svimo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios. *arXiv preprint arXiv:2506.02444*, 2025. 3

[15] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024. 3

[16] Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Errui Ding, Yu Wu, and Jingdong Wang. Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17550–17560, 2025. 3

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023. 3, 5

[18] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023. 2

[19] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2093–2103, 2024. 2

[20] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17461–17474, 2025. 2, 3

[21] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, pages 1–12. Wiley Online Library, 2023. 3

[22] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 3

[23] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Liefeng Bo, and Di

Huang. I4vgen: Image as free stepping stone for text-to-video generation. *arXiv preprint arXiv:2406.02230*, 2024. 3

[24] 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. *arXiv preprint arXiv:2307.04725*, 2023. 3

[25] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 3

[26] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 3

[27] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[28] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *Advances in Neural Information Processing Systems*, 35:23805–23817, 2022. 3

[29] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2005–2015, 2025. 5

[30] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7136–7146, 2025. 3

[31] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021. 3

[32] Shijian Jiang, Qi Ye, Rengan Xie, Yuchi Huo, and Jiming Chen. Hand-held object reconstruction from rgb video with dynamic interaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12220–12230, 2025. 2

[33] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 1

[34] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 5

[35] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[36] 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. 3

[37] Patrick Kwon, Chen Chen, and Hanbyul Joo. Graspdiffusion: Synthesizing realistic whole-body hand-object interaction. *arXiv preprint arXiv:2410.13911*, 2024. 3

[38] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *arXiv preprint arXiv:2503.00779*, 2025. 2

[39] Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, et al. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the computer vision and pattern recognition conference*, pages 11971–11981, 2025. 3

[40] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025. 2, 3

[41] Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17416–17425, 2025. 3

[42] Wenyi Li, Haoran Xu, Guiyu Zhang, Huan-ang Gao, Mingju Gao, Mengyu Wang, and Hao Zhao. Fairdiff: Fair segmentation with point-image diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 617–628. Springer, 2024. 3

[43] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. *arXiv preprint arXiv:2402.14810*, 2024. 3

[44] Xingyu Liu, Pengfei Ren, Qi Qi, Haifeng Sun, Zirui Zhuang, Jing Wang, Jianxin Liao, and Jingyu Wang. Generalizable hand-object modeling from monocular rgb images via 3d gaussians. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2

[45] Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references. *arXiv preprint arXiv:2502.09614*, 2025. 2

[46] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 3

[47] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 3

[48] Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, and Yebin Liu.

Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12209–12219, 2025. 2, 3, 5, 6

[49] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 3, 5

[50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4

[51] Anton Pelykh, Ozge Mercanoglu Sincan, and Richard Bowden. Giving a hand to diffusion models: a two-stage approach to improving conditional human image generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 3

[52] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12242–12254, 2025. 2

[53] Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7026–7036, 2025. 3

[54] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2024. 3

[55] Pengfei Ren, Jingyu Wang, Haifeng Sun, Qi Qi, Xingyu Liu, Menghao Zhang, Lei Zhang, Jing Wang, and Jianxin Liao. Prior-aware dynamic temporal modeling framework for sequential 3d hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6476–6487, 2025. 2

[56] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3, 6

[57] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. Maskhand: Generative masked modeling for robust hand mesh reconstruction in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8372–8383, 2025. 2

[58] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[59] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022. 5

[60] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *European Conference on Computer Vision*, pages 414–430. Springer, 2024. 2, 3, 5, 6

[61] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 5

[62] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[63] Benzhi Wang, Jingkai Zhou, Jingqi Bai, Yang Yang, Weihua Chen, Fan Wang, and Zhen Lei. Realishuman: A two-stage approach for refining malformed human parts in generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7509–7517, 2025. 3

[64] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. *arXiv preprint arXiv:2406.06843*, 2024. 2

[65] Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. Magichoi: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5957–5968, 2025. 2

[66] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023. 3

[67] Ruben Wiersma, Julien Philip, Miloš Hašan, Krishna Mullia, Fujun Luan, Elmar Eisemann, and Valentin Deschaintre. Uncertainty for svbrdf acquisition using frequency analysis. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 3

[68] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023. 3

[69] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20953–20962, 2022. 2, 3

[70] 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

[71] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 5

[72] Kai Ye, Yuhang Wu, Shuyuan Hu, Junliang Li, Meng Liu, Yongquan Chen, and Rui Huang. \textsc {Gen2Real}: Towards demo-free dexterous manipulation by harnessing generated video. *arXiv preprint arXiv:2509.14178*, 2025. 3

[73] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489, 2023. 3

[74] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7072–7084, 2025. 2

[75] Zhenjun Yu, Wenqiang Xu, Pengfei Xie, Yutong Li, Brian W Anthony, Zhuorui Zhang, and Cewu Lu. Dynamic reconstruction of hand-object interaction with distributed force-aware contact representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8590–8599, 2025. 2

[76] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024. 2, 3, 5

[77] Guiyu Zhang, Huan-ang Gao, Zijian Jiang, Hao Zhao, and Zhedong Zheng. Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling. *arXiv preprint arXiv:2410.11236*, 2024. 3

[78] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024. 2, 3, 6

[79] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1805–1815, 2025. 2

[80] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025. 3

[81] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3, 4, 8

[82] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8521–8531, 2024. 2, 3, 4

[83] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[84] Yufei Zhang, Zijun Cui, Jeffrey O Kephart, and Qiang Ji. Diffusion-based 3d hand motion recovery with intuitive physics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7306–7317, 2025. 2

[85] Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. *arXiv preprint arXiv:2505.18947*, 2025. 3

[86] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10–18, 2017. 2

[87] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020. 3

[88] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27683–27693, 2025. 3

[89] Zhuoran Zhao, Linlin Yang, Pengzhan Sun, Pan Hui, and Angela Yao. Analyzing the synthetic-to-real domain gap in 3d hand pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12255–12265, 2025. 2

[90] Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22584–22594, 2025. 3

[91] Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. Megohand: Multimodal egocentric hand-object interaction motion generation. *arXiv preprint arXiv:2505.16602*, 2025.

[92] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2024. 2, 3, 6, 8