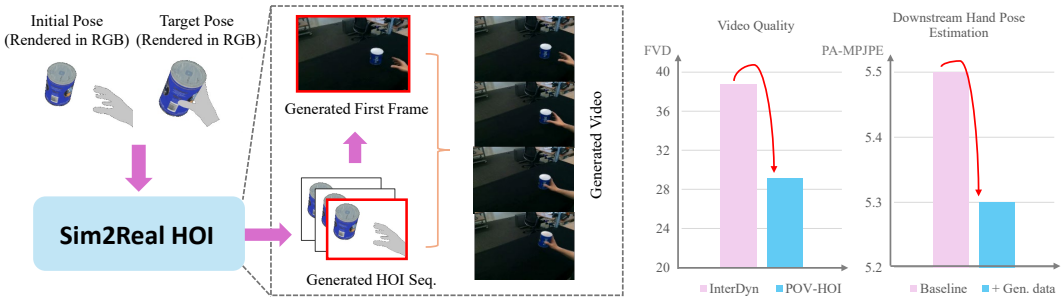
SIM2REAL-HOI: SIM-TO-REAL HOI VIDEO GENERATION VIA DECOUPLED MOTION–APPEARANCE DIFFUSION

Anonymous authors
Paper under double-blind review



(a) Overview of proposed Sim2Real-HOI

(b) Quantitative comparison on video quality and downstream task

Figure 1: (a) **Overview of proposed Sim2Real-HOI**. Given the initial and desired final poses of both the hand and object, along with the object mesh, Sim2Real-HOI generates photo-realistic in-the-wild videos by decoupling motion and appearance through a diffusion-based generative process, thereby closing the sim-to-real gap without paired real-world supervision. (b) **Evaluation on video quality and downstream task.** Our experiments reveal that Sim2Real-HOI not only generates videos with superior perceptual quality, as evidenced by a lower Fréchet Video Distance (FVD) (Unterthiner et al., 2018) against baselines, but also that these videos serve as effective synthetic data. Incorporating them into training reduces the error of a downstream hand pose estimator, outperforming the model trained solely on real videos.

ABSTRACT

We present Sim2Real-HOI, a zero-shot framework that closes the sim-to-real gap for hand–object interaction (HOI) video generation using the initial and target poses of both hand and object. Controllable diffusion models like InterDyn and ManiVideo stumble at scale when moving simulation to reality: the quality of generated videos are suboptimal, and they rely on simulator-unobtainable cues such as the first frame. Sim2Real-HOI addresses the problem in two stages: (1) an appearance generator that models both appearance and background using a controllable image diffusion model, and (2) a motion transfer model that transfers motion, generated by a pretrained hand pose generator, to real-world video through a controllable video diffusion model. To improve performance, we incorporate multiple types of conditions that ensure the generated output aligns with the geometry, semantics, and fine details of the hand pose. Extensive experiments on DexYCB and OAKINK2 demonstrate that Sim2Real-HOI enhances the generated quality compared to the best prior work and results in a lower error rate when the generated videos are used to train downstream hand-pose estimators. The code and pre-trained weights will be made publicly available.

1 INTRODUCTION

The ability to manipulate objects with two hands represents a fundamental human skill, and the computational understanding of this capability—referred to as Hand-Object Interaction (HOI) un-

Method \ Input Req.	Background	HOI Pose Seq	Object Appearance	Hand Apperance
CosHand	✓	✓	✓	✓
InterDyn	✓	✓	✓	✓
ManiVideo	✓	✓	✓	✓
Sim2Real-HOI (Ours)	✗	✓*	✗	✗

Table 1: Comparison of input requirements for HOI video generation. A checkmark (✓) signifies a condition provided as input to the model, whereas a cross (✗) indicates a condition that is not provided and will be synthesized. The symbol (*) denotes a setting where only the initial and target states are specified.

derstanding—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 (Zhou et al., 2024; Pavlakos et al., 2024; Dong et al., 2024) and enabling realistic human-to-robot motion transfer (Liu et al., 2025; Lepert et al., 2025; Li et al., 2025). The critical challenge, however, lies in the data itself. Despite considerable investments in collecting real-world HOI sequences with detailed annotations (Liu et al., 2022; Fu et al., 2025; Yang et al., 2022), the reliance on costly and labor-intensive manual labeling poses a fundamental limitation to scalability.

The advent of video diffusion models promises scalable generation of HOI videos. However, state-of-the-art methods (Pang et al., 2025; Akkerman et al., 2025; Sudhakar et al., 2024) critically depend on being conditioned on the first frame of the video, which creates a two-fold problem: (1) a significant *input bottleneck*, as obtaining a first frame that is geometrically consistent with the provided initial hand-object pose sequence is challenging, and (2) a *diversity bottleneck*, as fixing the first frame severely limits the potential for visual randomization, which is essential for data augmentation. Overcoming these bottlenecks by generating realistic videos from minimal inputs constitutes a key unsolved problem.

In this paper, we introduce a pioneering sim-to-real HOI video generation framework that requires *only* the initial and target poses, along with object geometry, as input. By integrating a novel decoupled 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. As demonstrated in Table 1 and Figure 1, our approach surpasses existing methods by jointly generating realistic foreground, background, and dynamically interpolated poses.

Our framework comprises two core stages. The **appearance generation** stage synthesizes a realistic initial frame using the controllable image diffusion model Flux (black-forest labs, 2024). This model is conditioned on a fusion of depth maps, semantic masks, and hand keypoint maps, which collectively ensure geometric accuracy, semantic coherence, and the preservation of fine-grained hand details. The **motion generation** stage then animates this frame into a video sequence. We first generate a plausible hand motion trajectory using a pre-trained model, which is subsequently rendered by a controllable video diffusion model (based on CogVideoX (Yang et al., 2024)). Crucially, the video model is conditioned on the same multi-modal inputs to maintain consistency with the generated HOI pose sequence.

We evaluate our method on the DexYCB (Chao et al., 2021) and OAKINK2 (Zhan et al., 2024) 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 an HOI video generation framework 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 appearance and motion synthesis, leveraging multi-modal conditions to achieve superior realism 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 (Agarwal et al., 2023; Ghosh et al., 2023; Christen et al., 2024). Prevailing data-driven approaches rely on supervised learning from large-scale, well-annotated datasets (Grady et al., 2021; Jiang et al., 2021; Karunratanakul et al., 2020; Dong et al., 2024; Pavlakos et al., 2024; Christen et al., 2022; Liu & Yi, 2024; Li et al., 2025; Zhong et al., 2025; Zhou et al., 2024). However, the scalability of these methods is constrained by their dependence on costly and difficult-to-acquire data (Fan et al., 2023; Hampali et al., 2020; Liu et al., 2022; 2024; Fu et al., 2025; Yang et al., 2022; Zhan et al., 2024; Chao et al., 2021). To circumvent this limitation, reinforcement learning (RL) has emerged as a promising alternative. Methods like (Christen et al., 2024; Xu et al., 2023) generate reference grasps before synthesizing motions, while GraspXL (Zhang et al., 2024a) 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.

2.2 Controllable Video Generation

Recent breakthroughs in video generation foundation models (Yang et al., 2024; Blattmann et al., 2023; Wan et al., 2025; Kong et al., 2024; Agarwal et al., 2025) have intensified interest in controllable generation that precisely aligns with user intent. While text-to-video and image-to-video models (Agarwal et al., 2025; Wan et al., 2025; Yang et al., 2024; Singer et al., 2022; Qing et al., 2024; Guo et al., 2023; Wiersma et al., 2025; Zhang & Agrawala, 2025) 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 (Zhang et al., 2023) and its variants (Gu et al., 2025; Guo et al., 2024b) enable conditioning on dense inputs, while works like VideoComposer (Wang et al., 2023) fuse multiple conditions for enhanced control. Camera motion has been explicitly modeled by embedding parameters into diffusion models (He et al., 2024; Bai et al., 2025). 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* (Hu et al., 2022; Kwon et al., 2024; Pelykh et al., 2024; Wang et al., 2025; Ye et al., 2023; Zhang et al., 2024b; Chen et al., 2025) typically conditions on 2D signals like segmentation masks and keypoints. However, these static methods cannot capture the dynamic nature of interactions. Recently, several studies (Sudhakar et al., 2024; Pang et al., 2025; Akkerman et al., 2025; Dang et al., 2025; Ye et al., 2025) have explored *HOI video generation*. InterDyn (Akkerman et al., 2025) conditions on hand mask sequences via ControlNet (Zhang et al., 2023), but under-utilizes the rich conditions available from simulators. ManiVideo (Pang et al., 2025) introduces an occlusion-aware representation but requires human appearance data, which is not available from simulators like GraspXL (Zhang et al., 2024a). 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.

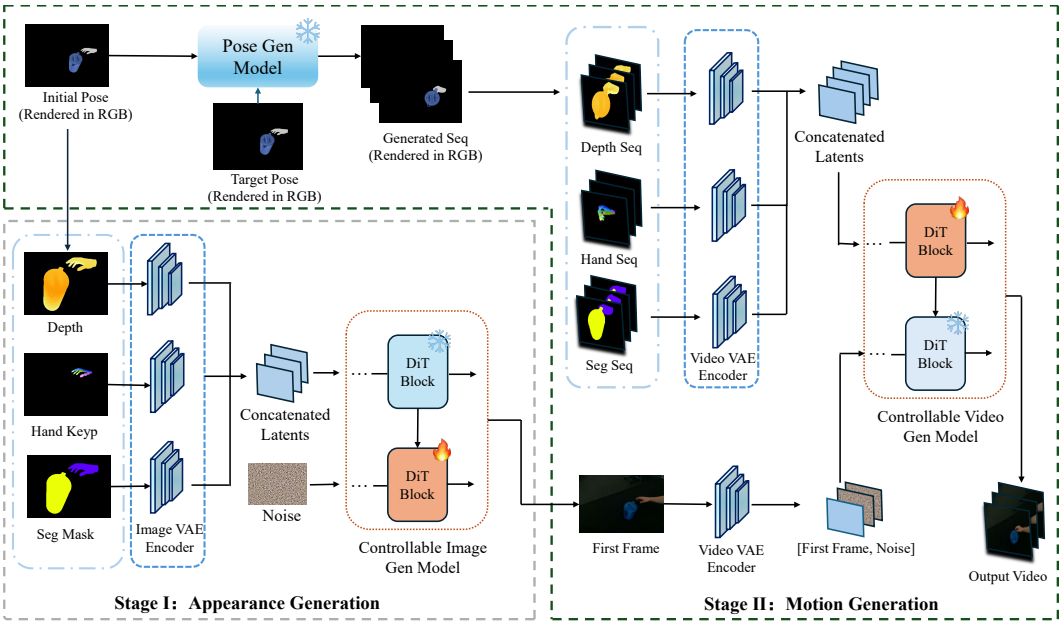


Figure 2: **Overview of our two-stage generation pipeline.** (1) **Appearance Generation:** A controllable image diffusion model synthesizes the first frame from multi-modal conditions (depth, semantic masks, keypoints). (2) **Motion Generation:** A pre-trained hand pose generator produces intermediate poses, which are then rendered into a full video sequence by a video diffusion model, conditioned on the same controls as appearance generation.

3 The Proposed Method

3.1 Overview

Figure 2 illustrates Sim2Real-HOI. Conditioned on an initial MANO (Romero et al., 2022) 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 modelling appearance and motion is notoriously hard because of the high-dimensional spatio-temporal manifold (Guo et al., 2024a). We therefore disentangle generation into two stages: (i) **appearance**—a pose-conditioned image diffusion model synthesises the first frame I_0 given initial hand–object poses and object mesh $(\mathbf{h}_0, \mathbf{o}_0, \mathbf{m})$ (Sec. 3.2); (ii) **motion**—a pretrained pose generator produces aligned sequences $\{\mathbf{h}_t, \mathbf{o}_t\}_{t=0}^{T}$, which are injected into a video diffusion model to animate I_0 into a photo-realistic clip (Sec. 3.3).

3.2 Stage I: 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 (Zhang et al., 2024b)—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 4.2 and 4.3, we explore the influence of every condition.

We fine-tune Flux (black-forest labs, 2024) with a ControlNet (Zhang et al., 2023) 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 (Peebles & Xie, 2023) 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 (Peebles & Xie, 2023) blocks, and f'_l is the output of the l-th layer of the duplicated DiT blocks whose input is the concentrated 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.

3.3 Stage II: Motion Generation Stage

To obtain the target video sequence we cascade a pretrained hand-motion generator with a controllable video diffusion model. As illustrated in Figure 2, GraspXL (Zhang et al., 2024a) consumes the initial MANO hand pose $\mathbf{h}_0$, the 6-DoF object pose $\mathbf{o}_0$ and object mesh $\mathbf{m}$ to produce aligned trajectories $\{\mathbf{h}_t, \mathbf{o}_t\}_{t=0}^{T}$. We rasterize depth maps, instance-level segmentation masks, and 2-D hand keypoint images at each frame. The pretrained video VAE encodes these conditions into a latent tensor of shape $\mathbb{R}^{\frac{T+1}{4} \times \frac{H}{8} \times \frac{W}{8} \times 16}$, after which we concentrate these latents and inject them into CogVideo-X through 12 duplicate DiT blocks, as outlined in Eq. 2. During training each cue is randomly masked with probability 0.2 to prevent over-reliance on any single modality.

4 Experiment

4.1 Experiment Settings

Datasets and Data Processing. We evaluate our method on two standard benchmarks for HOI video generation: DexYCB (Chao et al., 2021) and OAKINK2 (Zhan et al., 2024). 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 (Hu et al., 2025), 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) (Zhang et al., 2018) and Peak Signal-to-Noise Ratio (PSNR).
- **Spatio-temporal Coherence:** We adopt Fréchet Video Distance (FVD) (Unterthiner et al., 2018) to evaluate overall video realism, using the implementation from (Skorokhodov et al., 2022).
- **Motion Fidelity:** We use the Motion Fidelity (MF) metric (Yatim et al., 2024) to quantify dynamic accuracy. For each video, we sample 100 foreground points (on hands/objects), track them using CoTracker3 (Karaev et al., 2024), and compare the trajectories between generated and ground-truth videos. For a ground-truth tracklet $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$ and a generated tracklet $\tilde{\mathcal{T}} = \{\tilde{\tau}_1, \ldots, \tilde{\tau}_T\}$ 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:

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

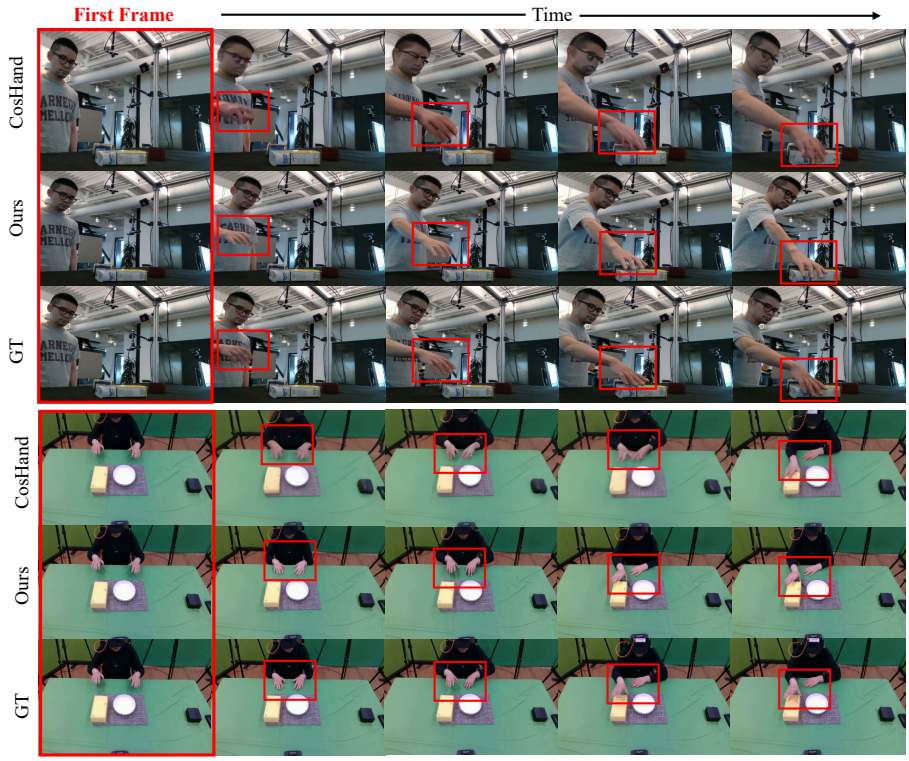


Figure 3: **Qualitative comparison against CosHand.** Example results on DexYCB and OAKINK2 highlight the strengths of our method in two key areas: (1) higher visual fidelity in both foreground and background generation, and (2) improved geometric accuracy of the synthesized hand poses.

Method	FVD (↓)	MF (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	MPJPE (↓)	Resolution
CosHand	58.51	0.591	0.139	0.767	23.20	30.05	256 x 256
InterDyn	38.83	0.680	0.119	0.848	24.86	-	256 x 384
ManiVideo	-	-	0.079	<u>0.913</u>	<u>30.10</u>	57.30	-
Ours w/ seg	33.23	0.695	0.077	0.900	29.27	21.14	480 x 720
Ours w/ depth	<u>30.00</u>	0.703	<u>0.070</u>	0.906	29.15	23.16	480 x 720
Ours w/ hand	33.41	**0.713**	0.086	0.901	29.07	<u>20.70</u>	480 x 720
Ours w/ all	**29.13**	<u>0.712</u>	**0.069**	**0.914**	**30.17**	**19.37**	480 x 720

Table 2: **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.

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 (Fan et al., 2023), measuring the average Euclidean distance between the 21 predicted and ground-truth hand joints after root alignment. Lower MPJPE indicates better pose estimation accuracy.

4.2 Main Results

Baselines. We compare our method against state-of-the-art HOI video generation approaches: ManiVideo (Pang et al., 2025), InterDyn (Akkerman et al., 2025), and CosHand (Sudhakar et al., 2024) on the DexYCB dataset (Chao et al., 2021). 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,

Method	FVD (↓)	MF (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	MPJPE (↓)
CosHand	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 3: **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.

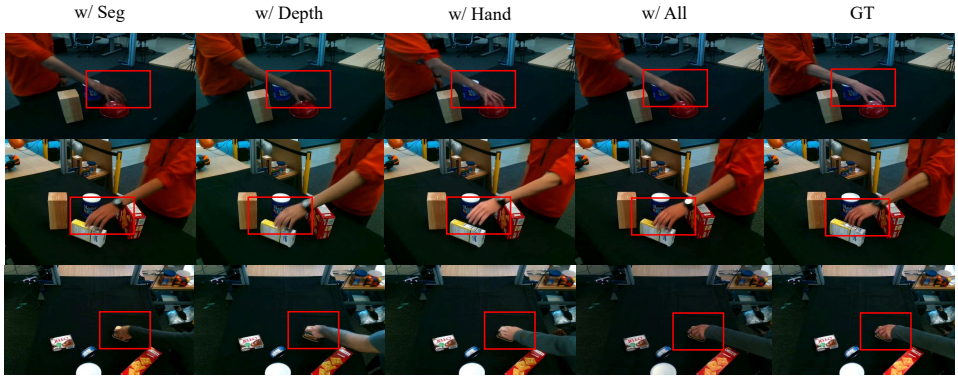


Figure 4: **Ablation study on input conditions on DexYCB dataset.**

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 2, 3) 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 3, 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 2, 3, and 4) yield three key observations:

Method	FVD (↓)	MF (↑)	LPIPS (↓)	SSIM (↑)	PSNR (↑)	MPJPE (↓)
Ours w/o seg	29.62	0.711	0.071	0.899	29.95	20.46
Ours w/o depth	29.53	0.711	0.073	0.902	29.57	19.92
Ours w/o hand	29.32	**0.712**	0.071	0.906	**30.60**	22.51
Ours w/ all	**29.13**	**0.712**	**0.069**	**0.914**	30.17	**19.37**

Table 4: **Ablation study on input conditions on DexYCB dataset.**

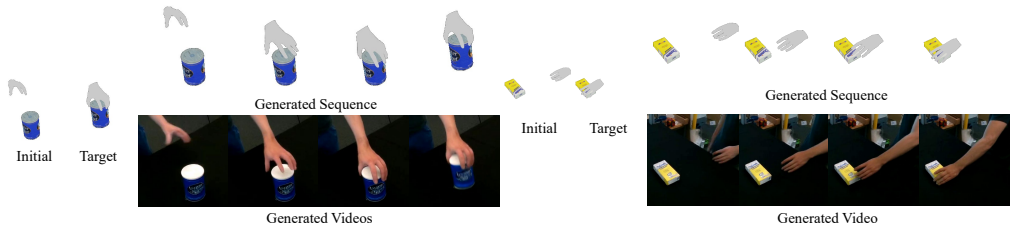


Figure 5: **Sim-to-real transfer results.** Sim2Real-HOI can generate realistic videos given initial and target states.

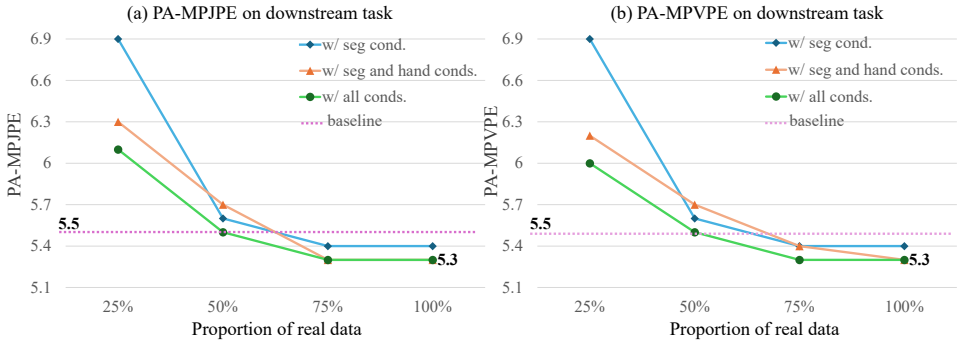


Figure 6: Data augmentation analysis with varying ratios of real data. We augment different portions of the DexYCB training set (25%, 50%, 75%, 100%) with our generated synthetic data. The **baseline** (dashed line) indicates performance when training solely on 100% of the real DexYCB data without synthetic augmentation.

- The performance improves with an increasing number of conditions, validating the effectiveness of our multi-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 4 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 Sim-to-Real Transfer

We conduct a sim-to-real transfer experiment to validate the effectiveness of our full pipeline. We employ GraspXL (Zhang et al., 2024a) as our hand motion generator due to its superior performance and strong generalization. Using objects from the DexYCB dataset, we randomly initialize the hand and object poses along with a target hand pose. GraspXL generates the intermediate motion sequence, which is then used to render the necessary conditions (depth, semantic masks, keypoints) for our video generation model. As shown in Figures 1 and 5, Our method effectively synthesizes realistic videos from minimal input, consisting solely of the initial and target poses, along with the object geometry. This capability stems from our decoupled generation architecture, which effectively integrates 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 (Zhou et al., 2024) as the pose estimation model, which regresses MANO parameters (Romero et al., 2022) from a single image. Our Sim2Real-HOI pipeline,

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 5: Downstream task evaluation on SimpleHand (Zhou et al., 2024).

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. PA-MPJPE/PA-MPVPE measure the average Euclidean distance (in mm) after Procrustes alignment between the predicted and ground-truth joints/vertices, respectively.

The quantitative results (Table 5) demonstrate that incorporating our generated data consistently improves hand pose estimation accuracy across all metrics. Figure 6 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.

4.6 Zero-Shot Results

To evaluate the generalizability of our approach, we test our model trained on the DexYCB dataset (single-hand interactions) directly on the OAKINK2 dataset (bimanual interactions) in a zero-shot setting. As shown in Figure 7, 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 (Zhang et al., 2023), which helps preserve the model's original generation quality while adapting to new conditioning signals.

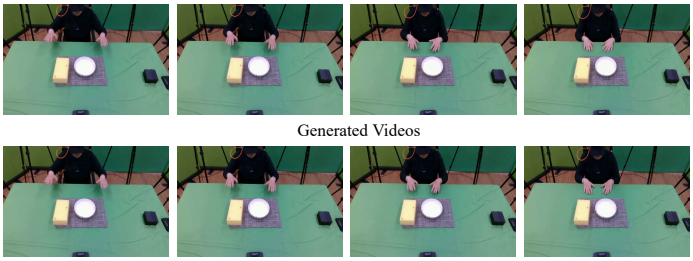


Figure 7: Zero-shot result on OAKINK2 dataset. We use the weight trained on DexYCB dataset.

5 Conclusion

This paper proposed Sim2Real-HOI, 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.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No human subjects or animal experimentation were involved in this study. All datasets used, including those provided by the authors, were sourced in compliance with relevant usage guidelines, ensuring the protection of privacy. We have taken measures to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. Our experimental setup is detailed in Section 4.1 and Section A.3. The code is available at https://anonymous.4open.science/r/Sim2Real-HOI-704C/.

Additionally, the public datasets used in this paper, such as DexYCB (Chao et al., 2021) and OAKINK2 (Zhan et al., 2024), are publicly available, ensuring consistent and reproducible evaluation results.

REFERENCES

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

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.

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*, pp. 12467–12479, 2025.

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.

black-forest labs. https://github.com/black-forest-labs/flux, 2024.

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.

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*, pp. 9044–9053, 2021.

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*, pp. 17448–17460, 2025.

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*, pp. 20577–20586, 2022.

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)*, pp. 3168–3175. IEEE, 2024.

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.

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.

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*, pp. 12943–12954, 2023.

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*, pp. 17461–17474, 2025.

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*, volume 42, pp. 1–12. Wiley Online Library, 2023.

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1471–1481, 2021.

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*, pp. 1–12, 2025.

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*, 2024a.

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.

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*, pp. 330–348. Springer, 2024b.

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*, pp. 3196–3206, 2020.

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.

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

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*, pp. 2005–2015, 2025.

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*, pp. 11107–11116, 2021.

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.

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)*, pp. 333–344. IEEE, 2020.

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.

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

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

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*, pp. 6991–7003, 2025.

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

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.

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*, pp. 21740–21751, 2024.

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*, pp. 21013–21022, 2022.

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*, pp. 12209–12219, 2025.

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*, pp. 9826–9836, 2024.

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

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)*, pp. 1–10. IEEE, 2024.

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*, pp. 6635–6645, 2024.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.

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

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.

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*, pp. 3626–3636, 2022.

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*, pp. 414–430. Springer, 2024.

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. URL `https://api.semanticscholar.org/CorpusID:54458806`.

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.

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*, volume 39, pp. 7509–7517, 2025.

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.

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*, pp. 1–12, 2025.

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*, pp. 4737–4746, 2023.

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*, pp. 20953–20962, 2022.

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.

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*, pp. 8466–8476, 2024.

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.

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*, pp. 22479–22489, 2023.

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*, pp. 445–456, 2024.

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*, pp. 386–403. Springer, 2024a.

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

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*, pp. 3836–3847, 2023.

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*, pp. 8521–8531, 2024b.

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*, pp. 586–595, 2018.

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*, pp. 22584–22594, 2025.

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*, pp. 1367–1376, 2024.

A APPENDIX

A.1 USE OF LLMS

We employ GPT-5 to improve and elevate the quality of our written content, primarily for enhancing accuracy and ensuring native-level expression. Its application is focused solely on refining language, rather than idea generation or other functions.

A.2 TOOLKIT

For access to our fully anonymous code toolkit, please visit: https://anonymous.4open.science/r/Sim2Real-HOI-704C/

A.3 IMPLEMENTATION DETAILS

Training Details: Our model was trained on a setup consisting 8 x NVIDIA 800 GPUs, with a batch size of 4 x 8 and a learning rate of 1×10^{-4}. The training process involved 8,000 training steps, using the AdamW optimizer and the DeepSpeed training architecture (Rajbhandari et al., 2020).

Evaluation Details: For the evaluation of video generation, we sample 1,600 videos, each consisting of 49 frames, from the test set. For the evaluation of Mean Per Joint Position Error (MPJPE), we utilize Hamer (Pavlakos et al., 2024) to estimate the hand joints in the generated videos, and compute the loss by comparing the estimated joint positions with the ground truth hand joints. To assess the performance on downstream tasks, we train the SimpleHand model for 200 epochs using its official implementation.

A.4 RESULTS

We provide more qualitative results in Figure 8 and Figure 9 for DexYCB dataset, Figure 10 and Figure 11 for OANINK2 dataset and Figure 12 for sim-to-real transfer.

Figure 8: More qualitative results on DexYCB dataset (a).

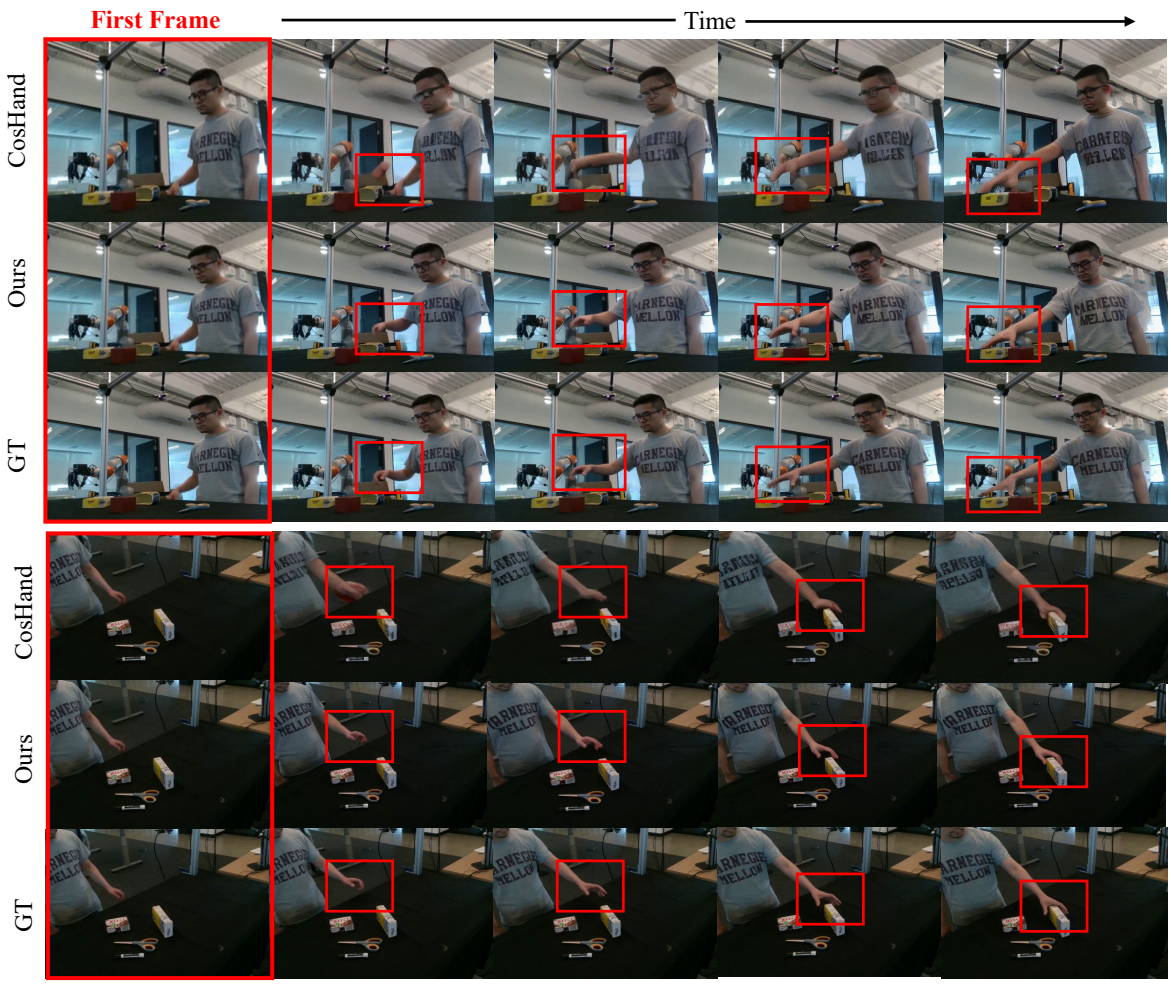


Figure 9: More qualitative results on DexYCB dataset (b).

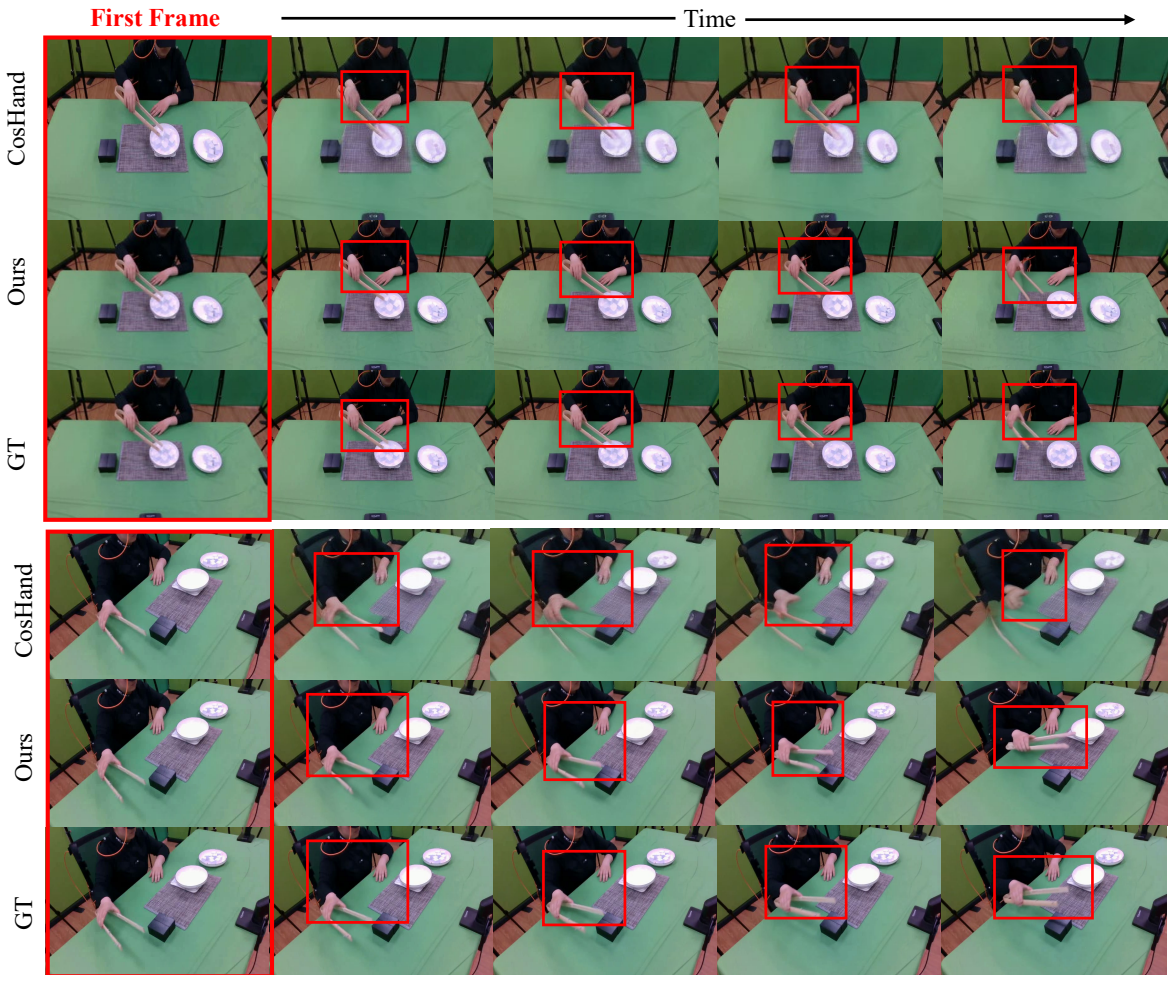


Figure 10: More qualitative results on OAKINK2 dataset (a).

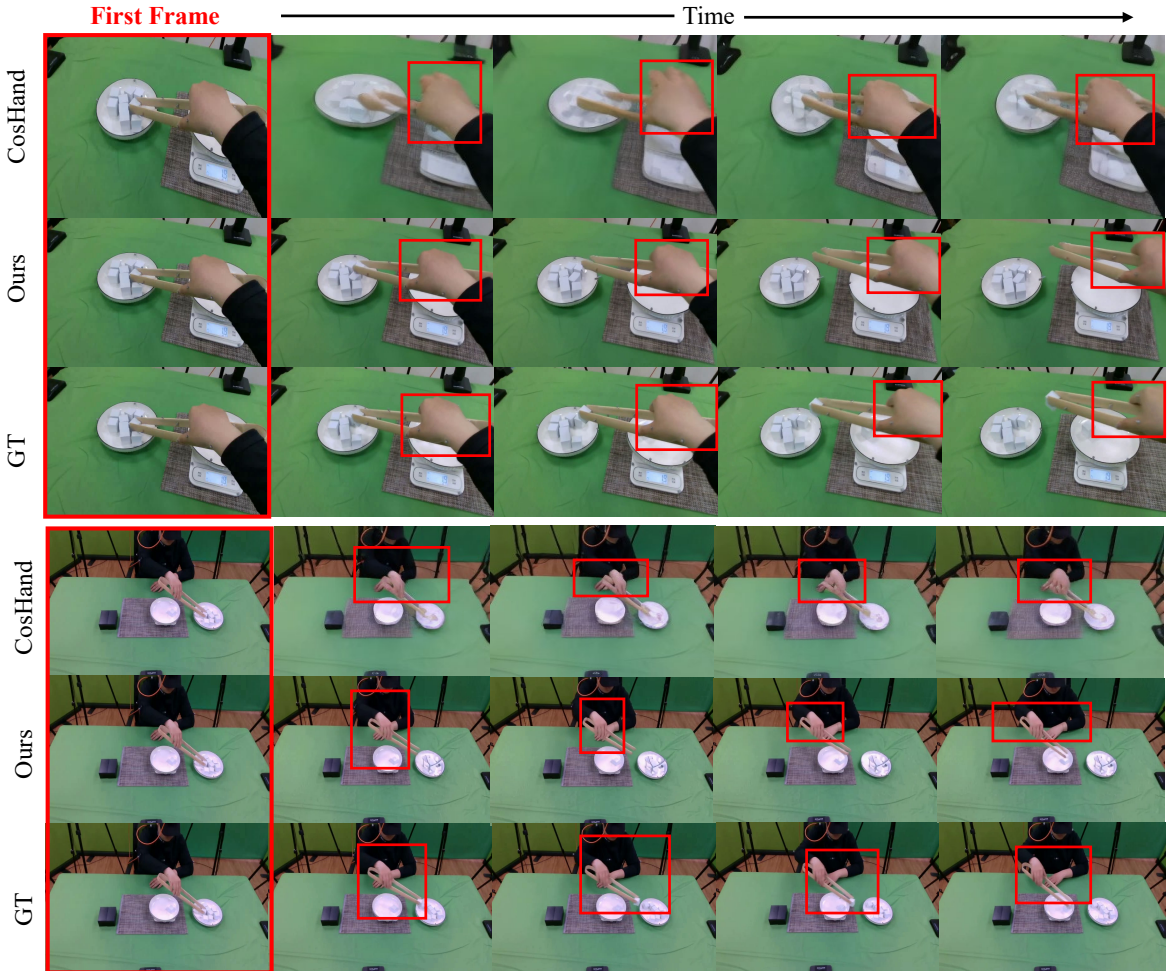


Figure 11: More qualitative results on OAKINK2 dataset (b).

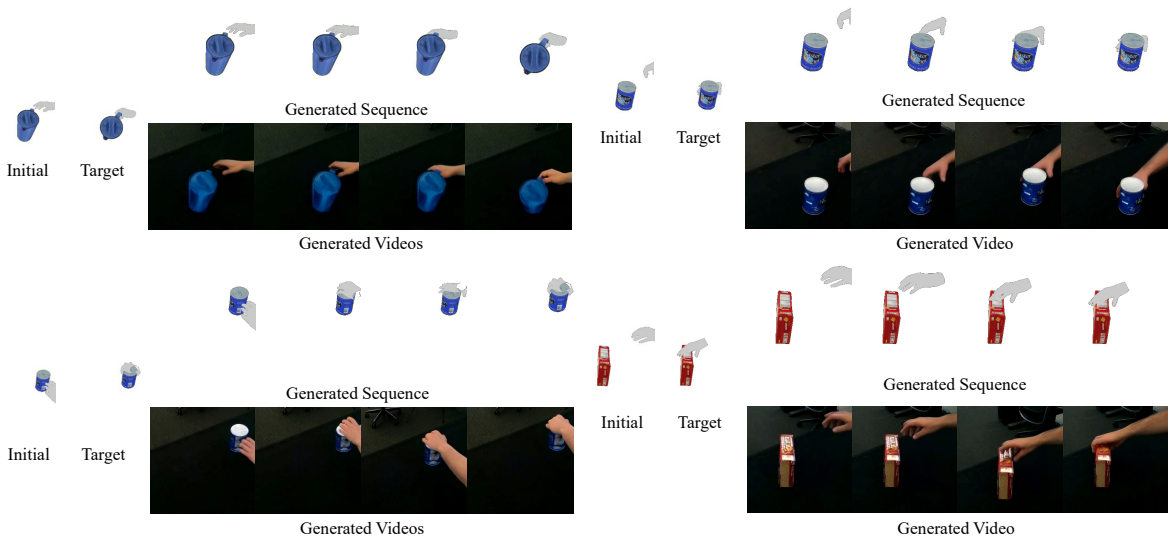


Figure 12: More Sim-to-real transfer results.