MVHOI: Bridge Multi-View Condition to Complex Human-Object Interaction Video Reenactment via 3D Foundation Model

Jinguang Tong[1,2,3*], Jinbo Wu[3*], Kaisiyuan Wang[3], Zhelun Shen[3], Xuan Huang[4], Mochu Xiang[3], Xuesong Li[1], Yingying Li[3], Haocheng Feng[3], Chen Zhao[3], Hang Zhou[3], Wei He[3], Chuong Nguyen[2], Jingdong Wang[3], Hongdong Li[1],

[1]The Australian National University [2]CSIRO [3]Baidu Inc. [4]Sun Yat-sen University

Abstract

Human–Object Interaction (HOI) video reenactment aims to transfer the interaction dynamics of a source video to a novel target object while preserving realistic hand–object coordination. Existing methods typically rely on sparse 2D motion controls and monocular references, which are insufficient for complex out-of-plane motion and large viewpoint changes. We present **MVHOI**, a two-stage framework combining implicit motion extraction, 3D-aware multi-view reasoning, and video generation. In the first stage, a motion extractor encodes object dynamics into implicit motion descriptors. Conditioned on these descriptors, our Motion-Driven Object Prior (MDOP) module queries a 3D foundation model over multi-view references of the target object and autoregressively predicts coarse object anchors, a sequence of images that track the object's evolving orientation and appearance under the source motion without any explicit pose estimation. In the second stage, a DiT-based video generation model uses these anchors as structural guidance and the multi-view references as appearance guidance. We further reuse cross-view attention from MDOP as a soft attention bias to reduce reference-view confusion. For long videos, a cross-iterative inference strategy refreshes subsequent object priors using refined video outputs. Experiments demonstrate consistent improvements over state-of-the-art methods in object fidelity, motion consistency, visual quality, and interaction realism.

1 Introduction

Reference-based human–object interaction (HOI) video reenactment aims to replace the interacted object in a source video with a novel target object while preserving the person, scene, and interaction dynamics (Fig. 1), enabling applications such as product demonstration and virtual advertising. Despite the strong generative capacity of recent Video Foundation Models (VFMs) (Wan et al. 2025; Kong et al. 2024; Yang et al. 2025), text or appearance conditions alone do not specify how the target object should evolve throughout the interaction, and reference-conditioned variants (Hu et al. 2025; Jiang et al. 2025; Liu et al. 2025; Zhang et al. 2025b) still fail to preserve the object's size and appearance under precise manipulation.

Many reference-based HOI methods (Fan et al. 2025; Huang et al. 2025; Xue et al. 2024; Wang et al. 2025b) describe object motion with sparse 2D control signals (e.g., bounding boxes or keypoints) and represent the target object with a monocular reference image. While effective for image-plane motion, these conditions become ambiguous under non-planar motion and large viewpoint changes: **(1) Ambiguous motion representation:** sparse 2D signals cannot determine the object's evolving view-dependent state, leaving the video model to infer complex 3D motion from insufficient evidence; and **(2) Ambiguous multi-view appearance conditioning:** a monocular reference provides incomplete appearance for unseen viewpoints, while naively supplying multiple references does not establish which view matches the object state in each frame. Consequently, the generated object may exhibit structural distortion or appearance drift as its orientation changes.

To address these challenges, we introduce MVHOI, a two-stage framework built on the spatial priors of 3D Foundation Models (3DFMs). Our core insight is that a 3DFM can consolidate sparse multi-view inputs into a unified latent object representation that jointly encodes geometry and appearance across viewpoints, allowing HOI reenactment to be reformulated as a motion-conditioned querying process over this representation.

In the first stage, our **Motion-Driven Object Prior (MDOP)** module encodes the temporal dynamics of the source object into implicit motion latents, which then query a 3DFM that consolidates the multi-view references into a unified target-object representation. This produces a coarse yet view-consistent target-object reenactment sequence that provides motion-aligned structural guidance for subsequent video generation, without explicit 6D pose estimation or offline tracking.

In the second stage, a video generation model synthesizes the high-fidelity HOI video conditioned on the coarse sequence and the multi-view references. Specifically, we reuse the internal attention responses produced by MDOP as a soft correspondence between the moving object and the reference views, and design an inference-time attention enhancement mechanism that modulates their contributions accordingly. The coarse reenactment sequence thus provides motion-aware structural guidance, while the reused attention cues improve appearance consistency across viewpoint changes,

*These authors contributed equally.

Figure 1: Given a source interaction video and multi-view images of a target object, MVHOI replaces the interacted object while faithfully following the source motion, preserving the target's identity and structure even under large rotations and viewpoint changes.

coupling the two stages through the same 3D-aware prior.

Our contributions are summarized as follows: **1)** We introduce MVHOI, a two-stage framework that connects implicit source-object motion extraction, 3D-aware multi-view fusion, and video generation for cross-object HOI reenactment under complex non-planar motion and large viewpoint changes. **2)** We propose the MDOP module, which encodes source-object transformations as implicit motion latents and transfers them to the target object via motion-queried multi-view fusion within a 3D foundation model, producing coarse object anchors that provide target-specific structural guidance without explicit 6D pose estimation. **3)** We reuse the cross-view attention responses of MDOP to coordinate multi-view appearance conditioning in the video generator and introduce a cross-iterative inference strategy to reduce structural and appearance drift in long-video generation.

2 Related Work

2.1 Controllable Video Generation and Motion Transfer

Early video generation methods (Wu et al. 2023; Blattmann et al. 2023; Guo et al. 2024) adapted image diffusion models (Rombach et al. 2022) to video, and transformer-based models (Brooks et al. 2024; Yang et al. 2025; Kong et al. 2024; Wan et al. 2025) scale up this paradigm with greatly improved temporal coherence and fidelity. Controllable variants preserve subject and object appearance through reference images or auxiliary conditioning branches (Hu et al. 2025; Jiang et al. 2025; Liu et al. 2025), while motion-transfer methods such as DisMo (Ressler-Antal et al. 2025) learn abstract motion representations that transfer across appearances and categories. However, these general-purpose methods do not associate source interaction dynamics with the multi-view geometry of a novel target object, and thus struggle to preserve object structure under substantial out-of-plane reorientation.

2.2 HOI Video Generation and Reenactment

Conditional HOI generation synthesizes interactions from predefined controls such as human poses, object trajectories, bounding boxes, meshes, or hand-object motion layouts (Xu et al. 2026a; Huang et al. 2025; Wang et al. 2025b; Pang et al. 2025), and geometry-aware approaches further inject spatial priors as explicit 3D representations (Chen et al. 2026) or hierarchical multi-view geometry and texture features (Xu et al. 2026b). These methods generate from predefined control signals rather than transferring interaction dynamics from a source video to a novel target object.

HOI reenactment methods instead replace the manipulated object within an existing interaction, via single-frame replacement with sequential warping (Xue et al. 2024), adaptive layout conditions (Fan et al. 2025), two-stage inpainting (Shen et al. 2025), or temporally balanced and spatially selective reference injection (Huang et al. 2026). However, none of these methods constructs a time-varying target-object prior that couples the source interaction motion with the target object's multi-view geometry.

Recent 3D foundation models (Wang et al. 2024, 2025a; Lin et al. 2026) learn transferable multi-view geometric representations, but are mostly used as static reconstruction or depth priors. In contrast, the 3D prior in MVHOI is motion-conditioned and dynamically aligned with the source interaction: source-motion latents and the target multi-view representation jointly produce a temporally varying object prior, whose geometry-aware attention cues additionally guide the video generation model.

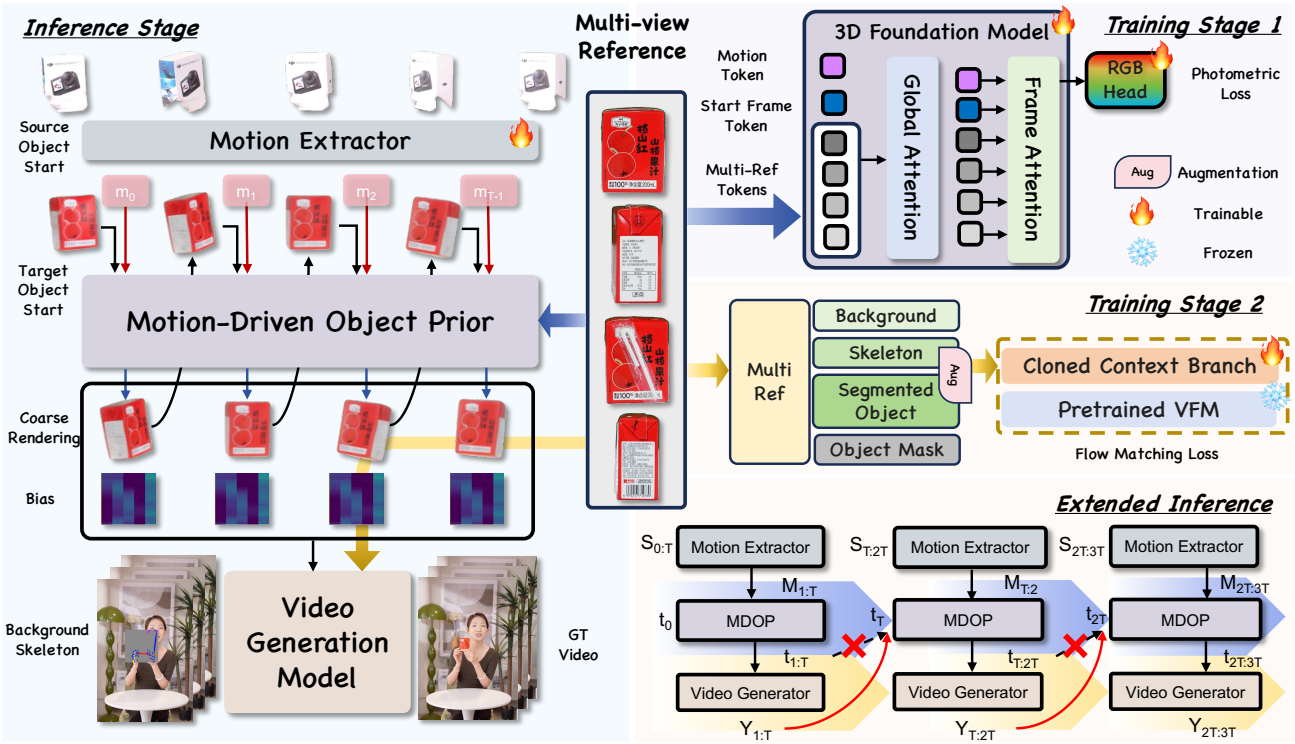

Figure 2: Overview of MVHOI. Given a source interaction video and multi-view images of a target object, the Motion-Driven Object Prior (MDOP) module converts implicit source-object dynamics into a sparse sequence of target-specific object anchors within a pretrained 3D-aware multi-view feature space. These anchors provide motion-aligned structural guidance to a DiT-based video generator, while the original multi-view references supply complementary appearance details. For long videos, the refined object state from one clip initializes the prior construction of the next clip.

3 Method

3.1 Overview

We address cross-object Human–Object Interaction (HOI) video reenactment. Given a source interaction video $\mathcal{V}_{src} = \{X_t\}_{t=1}^{T}$ and N reference images $\mathcal{I}_{ref} = \{I_i\}_{i=1}^{N}$ depicting a target object from different viewpoints, our goal is to synthesize $\mathcal{V}_{syn} = \{Y_t\}_{t=1}^{T}$, which should preserve the interaction dynamics of the source video while replacing the originally interacted object with the target object and maintaining the target's structure and appearance under substantial viewpoint changes.

As illustrated in Fig. 2, MVHOI follows a two-stage pipeline. In the first stage (Stage I), a motion extractor encodes the temporal dynamics of the source object into a sequence of implicit motion latents. These latents are fed into our **Motion-Driven Object Prior (MDOP)** module, which builds on a pretrained 3D foundation model to convert them into motion queries and jointly process the current target-object state and its multi-view references. MDOP then autoregressively predicts a sparse sequence of coarse target-object anchors that capture the evolving object state without explicit 6D pose estimation or 3D reconstruction. In the second stage (Stage II), a multi-reference video generation model uses the coarse anchors as structural guidance and the multi-view references as appearance guidance to synthesize the final HOI video with fine-grained object details. During long-video inference, the two components are further coupled across clips to reduce accumulated structural and appearance drift.

3.2 Motion-Driven Object Prior Construction

The objective of MDOP is to transfer the source-object motion to the target object and predict the evolving visual states of the target object. Instead of relying on the sparse explicit motion signals used in prior works (Xu et al. 2026a; Wang et al. 2025b), which become ambiguous under out-of-plane rotations, we use implicit motion descriptors extracted from the source-object sequence to condition the target-object state prediction.

Implicit source-motion encoding. We first extract the source-object crops $\mathcal{S} = \{S_t\}_{t=1}^{T}$ using corresponding object masks. We adapt the motion encoder $\mathcal{M}_\theta$ from DisMo (Ressler-Antal et al. 2025) to obtain a sequence of implicit motion latents $\mathbf{M} = \mathcal{M}_\theta(\mathcal{S}) = \{m_t\}$ with $m_t \in \mathbb{R}^{d_m}$, computed at a temporal stride of $\Delta = 4$ frames. Unlike explicit motion representations such as 6D poses or keypoints, m_t directly captures how the source object changes within a short video clip. This avoids the need for an additional pose estimation or pose fitting step.

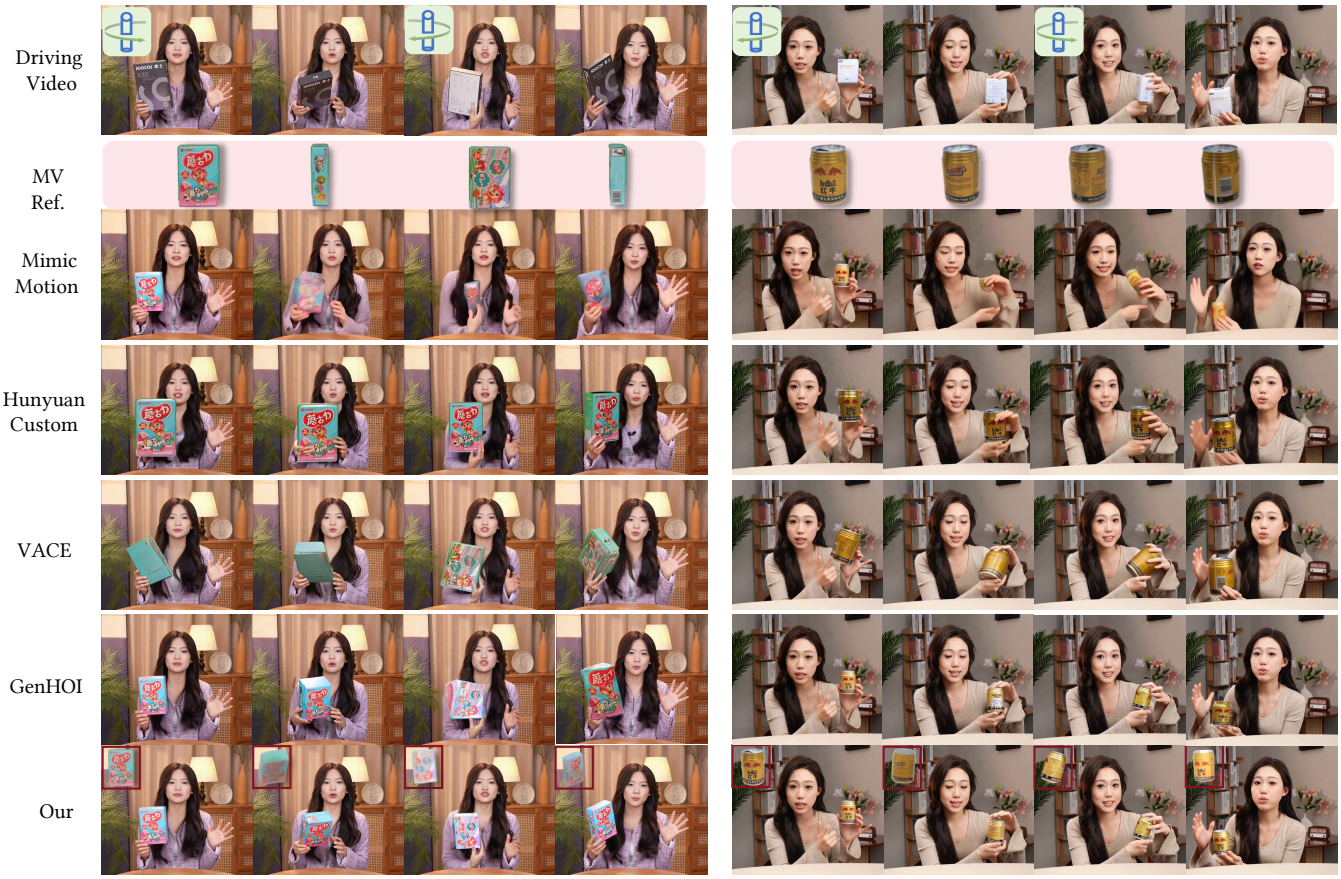

Figure 3: Qualitative comparison with state-of-the-art methods. Compared to MimicMotion (Zhang et al. 2025a), HunyuanCustom (Hu et al. 2025), VACE (Jiang et al. 2025), and GenHOI (Huang et al. 2026), our method generates object motion that more faithfully follows the driving video. Moreover, MVHOI better preserves the identity and structural consistency of the target object during complex manipulation. Detailed experimental settings are provided in Section 4.

Motion injection into the 3D foundation model. Although the implicit motion latent m_t captures the source-object dynamics from time t to $t+\Delta$, it does not directly specify how the target object should evolve under this transformation. We therefore introduce, inside the 3D foundation model, a learnable motion-condition feature adapter $\mathcal{A}_\phi$ that projects the source-motion latent into a motion query based on the current target-object state $\hat{O}_t \in \mathbb{R}^{3\times H\times W}$. An image encoder $\mathcal{E}$ extracts a spatial feature map from $\hat{O}_t$, a modulation layer $\mathcal{H}$ maps m_t to channel-wise scale and shift parameters $\gamma_t, \beta_t \in \mathbb{R}^C$, and a decoder $\mathcal{D}$ maps the modulated feature to the motion query:

$$P_t = \mathcal{D}\Big(\gamma_t \odot \mathrm{RMSNorm}\big(\mathcal{E}(\hat{O}_t)\big) + \beta_t\Big), \qquad (1)$$

with $[\gamma_t;\beta_t] = \mathcal{H}(m_t)$, where C is the feature-channel dimension, $[\cdot;\cdot]$ denotes channel-wise concatenation, $\odot$ denotes element-wise multiplication, and γ_t and β_t are broadcast over the spatial dimensions. The three components together constitute the learnable adapter $\mathcal{A}_\phi$; its architectural details are provided in the Appendix.

Importantly, $P_t \in \mathbb{R}^{3\times H\times W}$, which encodes the transition from t to $t+\Delta$, is neither an explicit pose map nor the predicted target-object frame at $t+\Delta$; it is patchified as an additional query view for the subsequent 3D-aware reasoning process described below.

3D-aware multi-view fusion and anchor decoding. Recent feed-forward 3D foundation models learn transferable cross-view representations from multiple visual observations (Wang et al. 2025a; Lin et al. 2026). We instantiate our multi-view fusion backbone $\mathcal{G}_\psi$ from Depth Anything 3 (Lin et al. 2026). At each sampled time step, it jointly processes the current target-object state, the N reference images, and the motion query:

$$Z_t = \mathcal{G}_\psi\left([\hat{O}_t, I_1, \ldots, I_N, P_t]\right). \qquad (2)$$

The backbone alternates intra-view and cross-view attention, enabling the motion query to aggregate the structure and appearance of the target object from its reference views. The resulting joint tokens serve as the internal representation for anchor prediction rather than an explicit reconstruction of the target object in 3D space.

We extract the features associated with the motion-query view, $Z_{t,P}$, and decode the next coarse object state with an RGB prediction head, $\hat{O}_{t+\Delta} = \mathcal{R}(Z_{t,P})$; in practice $\mathcal{R}$ consumes features from several intermediate layers of $\mathcal{G}_\psi$ (see Appendix). Starting from an initial target-object state $\hat{O}_0$, MDOP applies this process autoregressively to obtain a sparse sequence of coarse object anchors $\hat{\mathcal{O}} = \{\hat{O}_0, \hat{O}_\Delta, \hat{O}_{2\Delta}, \ldots\}$. Although these anchors do not preserve all high-frequency texture details, they encode the target object's evolving orientation, silhouette, and coarse appearance under the source motion.

3.3 Video Generation with Multi-Reference Conditioning

Given the sparse object anchors $\hat{\mathcal{O}}$ produced by MDOP, we employ a DiT-based video inpainting model (Wan et al. 2025) to synthesize the final HOI video. The coarse anchors and multi-view references provide complementary conditions: the anchors constrain the frame-dependent object state, whereas the reference images provide high-quality appearance information that is absent from the coarse predictions.

Coarse-prior and reference conditioning. To inject the two conditions into the video-generation backbone, we adopt an additional context branch following VACE (Jiang et al. 2025). The target-object reference images are prepended to the context sequence and processed by the cloned context branch, preserving their photometric information. Meanwhile, the temporally aligned coarse anchors are incorporated into the masked video stream using the HOI masks. Self-attention then propagates information among the noisy video tokens, coarse structural conditions, and multi-view reference tokens.

Directly optimizing this component with MDOP predictions would require repeated inference of the 3D foundation model throughout training, leading to prohibitive computational costs. We therefore construct proxy guidance from the ground-truth object crops. Naively using clean object crops, however, would cause appearance leakage and allow the video model to ignore the multi-view references. To approximate the degraded characteristics of MDOP predictions, we apply geometric perturbations, photometric jittering, blur, and noise to the proxy guidance. These augmentations suppress high-frequency appearance information while preserving coarse shape and orientation, encouraging the video model to use the object anchors for structural alignment and the reference images for detailed appearance synthesis.

Inference-time attention enhancement. Although multi-view reference tokens provide complementary appearance information, appearance-driven attention may assign excessive weight to reference views that are incompatible with the current object state. Inspired by prior studies showing that diffusion generation can be controlled through attention manipulation (Hertz et al. 2023), we use the cross-view associations formed inside MDOP to modulate reference conditioning during inference.

Let $\bar{A}_{t,i}$ denote the mean attention between the motion-query tokens and the tokens of the i-th reference view, averaged over all heads and token pairs at a designated layer of the MDOP fusion backbone $\mathcal{G}_\psi$. We normalize these associations into a frame-to-view weight and convert it into a reference-attention bias:

$$w_{t,i} = \frac{\bar{A}_{t,i}}{\sum_{i'=1}^{N} \bar{A}_{t,i'}}, \qquad B_{t,i} = \alpha \log \frac{w_{t,i} + \epsilon}{1 - w_{t,i} + \epsilon}, \quad (3)$$

where α controls the bias strength and ϵ ensures numerical stability. Rather than treating $w_{t,i}$ as a hard retrieval decision, we use it to softly modulate the relative influence of different reference views: the bias is broadcast to the reference tokens associated with view i and added to the corresponding attention logits:

$$\mathrm{Attn}(Q, K, V; B) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + B\right) V. \quad (4)$$

This operation increases the contribution of reference tokens whose views are more compatible with the current coarse object state. We therefore refer to it as *inference-time attention enhancement* (AE), rather than explicit appearance retrieval. The extracted associations shift smoothly across reference views as the object orientation evolves (visualized in the Appendix), indicating that MDOP encodes a meaningful view-selection signal.

3.4 Training Objectives

As discussed above, we train MDOP and the multi-reference video generator separately for computational efficiency, and integrate them only at inference time.

Motion-driven object prior construction. We train MDOP to predict coarse target-object anchors that preserve the source object's motion dynamics while maintaining the target object's appearance. These predicted anchors are supervised by

$$\mathcal{L}_{\mathrm{MDOP}} = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_{\mathrm{LPIPS}} + \lambda_3 \mathcal{L}_{\mathrm{SSIM}}, \quad (5)$$

combining pixel reconstruction, perceptual similarity, and structural similarity terms.

Multi-reference video generation. We train the video generator with the standard flow-matching objective (Lipman et al. 2023; Liu, Gong, and Liu 2023),

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\big[\|v_\vartheta(x_t, t, c) - u_t\|_2^2\big], \quad (6)$$

where v_ϑ is the predicted velocity field, u_t is the target velocity, and c denotes the conditioning inputs. Following the HOI-region emphasis used in AnchorCrafter (Xu et al. 2026a), we assign a larger weight to the relatively small interaction region. Let $\mathbf{M}_{\mathrm{HOI}} \in \{0, 1\}$ denote its binary mask. We define the spatial weighting function as

$$\rho(j) = 1 + \left(\eta \frac{S}{S_{\mathrm{HOI}}} - 1\right) \mathbf{M}_{\mathrm{HOI}}(j), \quad (7)$$

where S and S_{HOI} are the areas of the full image and the HOI region, respectively. The final video objective is

$$\mathcal{L}_{\mathrm{video}} = \mathbb{E}\left[\frac{1}{S}\sum_j \rho(j)\, \|v_\vartheta(x_t, t, c)_j - u_t(j)\|_2^2\right], \quad (8)$$

Method	Self-Reenactment						Cross-Reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	O-CLIP ↑	FID ↓	FVD ↓	O-CLIP ↑
MimicMotion	19.34	0.806	0.233	56.13	632.9	0.685	84.52	774.2	0.470
VACE	27.25	<u>0.954</u>	0.036	47.43	204.6	0.706	62.54	308.1	0.556
HuMo	8.71	0.497	0.716	231.03	2120.8	0.449	225.81	2282.5	0.322
HunyuanCustom	25.56	0.925	0.090	48.88	247.4	0.791	62.68	349.7	0.640
GenHOI	<u>29.60</u>	0.949	<u>0.035</u>	<u>19.24</u>	**76.8**	**0.866**	<u>48.01</u>	<u>280.4</u>	**0.680**
Ours	**30.21**	**0.960**	**0.025**	**17.90**	<u>86.6</u>	<u>0.848</u>	**41.14**	**274.0**	<u>0.645</u>

Table 1: Quantitative comparison with baseline methods on Self-Reenactment and Cross-Reenactment. Our method achieves the best results on most metrics, with clear advantages in reconstruction fidelity, perceptual quality, and temporal consistency. Best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

where the factor S/S_{HOI} compensates for the small spatial extent of the interaction region, and η controls its relative importance.

3.5 Cross-Iterative Long-Video Inference

Long-video generation requires maintaining target-object identity and motion stability over multiple temporal segments. Independently rolling out coarse anchors over the entire sequence can accumulate structural errors, whereas continuously extending the video generator from its previous predictions may cause appearance drift.

We therefore couple the two components across successive clips. For temporal segment s, MDOP first predicts a set of coarse object anchors $\hat{\mathcal{O}}^{(s)}$. The video generator then synthesizes a refined clip $\mathcal{Y}^{(s)}$ using the anchors and target-object references. Instead of initializing the next MDOP rollout from its previous coarse prediction, we extract the target-object state from the final refined frame of $\mathcal{Y}^{(s)}$, $\hat{O}_0^{(s+1)} = \mathrm{Crop}_{obj}(Y_{\mathrm{end}}^{(s)})$. This cross-iterative loop lets MDOP repeatedly provide motion-aligned structural constraints to the video generator, while the refined video output refreshes the visual state used by MDOP, mitigating the accumulation of coarse-prediction errors and long-term appearance drift.

4 Experiments

4.1 Implementation Details

We initialize the implicit motion extractor from the pretrained DisMo (Ressler-Antal et al. 2025). For the MDOP module, we construct it from a pretrained 3D foundation model (Lin et al. 2026) with the proposed motion-condition adapter and train on a synthetic cross-object dataset rendered from Objaverse (Deitke et al. 2023) and a self-collected dataset ($\sim$100 hours) with multi-view reference images. As for the video generation model, we initialize it from the pretrained Wan2.1-T2V-14B (Wan et al. 2025) with the additional context branch and train it on the same self-collected dataset. More details about data collection, preprocessing, and training are provided in the Appendix.

4.2 Evaluation Setup

We evaluate MVHOI under two settings: **Self-Reenactment**, which reconstructs HOI videos from the masked source video and references depicting the same object, and **Cross-Reenactment**, which replaces the interacted object with a different one. All source and target pairs are fixed in advance and shared by all methods.

Metrics. We report FID (Heusel et al. 2017) and FVD (Unterthiner et al. 2018) to assess video quality, and Object-CLIP (O-CLIP) (Xu et al. 2026a; Huang et al. 2025) for target-object fidelity. For Self-Reenactment, where frame-aligned ground truth is available, we additionally report PSNR (Al Najjar 2024), SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018) for reconstruction fidelity. Cross-Reenactment is further evaluated with a human study of motion consistency (MC), visual quality (VQ), and HOI realism (HR); exact metric implementations and the study protocol are described in the Appendix.

Baselines. We compare MVHOI with MimicMotion (Zhang et al. 2025a), VACE (Jiang et al. 2025), HunyuanCustom (Hu et al. 2025), HuMo (Chen et al. 2025), and GenHOI (Huang et al. 2026), spanning pose-guided human animation, unified controllable video editing, multi-subject reference-conditioned generation, and HOI reenactment. All methods are evaluated with their official checkpoints, except VACE, which we fine-tune on our collected dataset with multiple reference views for a fair comparison. Per-method input adaptations are detailed in the Appendix.

4.3 Experimental Results

Evaluation of motion-driven object priors. We first evaluate MDOP in isolation from the video generator on a cross-object transfer task, where both MDOP and DisMo (Ressler-Antal et al. 2025) receive the same pretrained motion representation, source motion, and target-object initialization; any gain is therefore attributable to target-aware multi-view fusion. As shown in Tab. 2, MDOP outperforms DisMo in both reconstruction fidelity and perceptual quality, and the qualitative results in Fig. 4 show correspondingly better structural and appearance consistency under viewpoint changes. These results support our central design: coupling implicit source motion with multi-view observations of the target object yields a more target-specific and structurally stable dynamic prior.

Figure 4: Qualitative comparison of cross-object motion transfer. The source and target are different objects. While DisMo (Ressler-Antal et al. 2025) transfers motion from a single target-object initialization, MDOP additionally reasons over multi-view target-object observations, leading to improved structural and appearance consistency under viewpoint changes.

Method	PSNR ↑	SSIM ↑	FID ↓	KID (10^{-3}) ↓
DisMo	16.56	0.8081	111.96	24.81
Ours	**20.75**	**0.8304**	**84.58**	**9.67**

Table 2: Quantitative comparison with DisMo on held-out cross-object pairs from Objaverse. MDOP improves both frame-aligned reconstruction fidelity and perceptual distribution quality.

Full HOI video reenactment. We evaluate the complete MVHOI framework under Self-Reenactment and Cross-Reenactment, and report quantitative results in Tab. 1. MVHOI achieves the best scores on most metrics across both settings: under Self-Reenactment it attains the best reconstruction fidelity and perceptual quality, accurately recovering real HOI videos from multi-view references of the same product, while under the more challenging Cross-Reenactment setting it obtains the best FID and FVD, indicating improved preservation of the source dynamics and stronger overall video quality. GenHOI (Huang et al. 2026) obtains higher O-CLIP scores, reflecting its strength in direct reference-appearance preservation, whereas MVHOI provides better distribution-level video quality and motion preservation by explicitly constraining the evolving object state through its motion-driven 3D-aware prior. The qualitative results in Fig. 3 support these numbers. MVHOI faithfully follows the source motion while preserving the target object's identity fidelity and multi-view appearance consistency under large movements, where the baselines often exhibit unstable object shape or drifting appearance.

User study. Table 3 reports the human evaluation on Cross-Reenactment for both short and long video generation. Our method receives clearly higher visual-quality ratings than all baselines, together with the best motion consistency and HOI realism. Moreover, MVHOI maintains its performance when moving from short to long video generation, whereas the ratings of competing methods degrade generally, demonstrating the effectiveness of our cross-iterative inference strategy for long-horizon stability. More qualitative video results are provided in the supplementary material.

Method	Short			Long		
	MC ↑	VQ ↑	HR ↑	MC ↑	VQ ↑	HR ↑
VACE	3.66	3.20	3.32	3.73	2.14	3.27
HunyuanCustom	3.59	3.29	3.11	3.49	3.56	3.17
GenHOI	4.20	4.10	4.15	4.07	3.94	3.89
Ours	**4.47**	**4.34**	**4.34**	**4.50**	**4.32**	**4.32**

Table 3: User study on Cross-Reenactment for short (single-clip) and long (10-second) video generation. MC, VQ, and HR denote motion consistency, visual quality, and HOI realism, rated on a 1–5 scale (higher is better).

Method	FID ↓	FVD ↓	O-CLIP ↑
Baseline	46.55	296.7	0.611
+MDOP	45.12	282.1	0.642
+MDOP+AE (full)	**41.14**	**274.0**	**0.645**

Table 4: Ablation study on Cross-Reenactment. MDOP denotes conditioning the video generator on the coarse object anchors from our Motion-Driven Object Prior module, and AE denotes the inference-time attention enhancement.

4.4 Ablation Study

We conduct an ablation study to evaluate the contribution of two key components in our framework: the **coarse object anchors produced by MDOP** and the **inference-time attention enhancement (AE)**. As the baseline, we train a multi-reference adapter conditioned solely on multi-view reference images. Quantitative results on Cross-Reenactment are reported in Tab. 4, with qualitative comparisons shown in the Appendix.

Overall, both components contribute consistently to improved generation quality and temporal stability. Conditioning on the MDOP anchors provides explicit motion-aware structural guidance, which stabilizes object dynamics and reduces temporal inconsistency. Building on this, AE further improves appearance fidelity by encouraging the model to retrieve viewpoint-consistent details from the multi-view references, thereby alleviating view confusion and enhancing cross-frame appearance consistency.

5 Conclusion

We present MVHOI, a framework that bridges multi-view object conditions and HOI video reenactment through 3D foundation models. The proposed Motion-Driven Object Prior (MDOP) module converts implicit source motion and multi-view references into motion-aligned structural guidance, an inference-time attention enhancement improves viewpoint-consistent appearance, and a cross-iterative inference strategy suppresses drift in long-video generation. Experiments across object reenactment, HOI video generation, and long-horizon synthesis validate explicit 3D-aware guidance coupled with multi-view appearance modeling, with consistent quantitative and perceptual gains.

References

Al Najjar, Y. 2024. Comparative analysis of image quality assessment metrics: MSE, PSNR, SSIM and FSIM. *International Journal of Science and Research (IJSR)*, 13(3): 110–114.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video Generation Models as World Simulators. OpenAI technical report.

Chen, L.; Ma, T.; Liu, J.; Li, B.; Chen, Z.; Liu, L.; He, X.; Li, G.; He, Q.; and Wu, Z. 2025. HuMo: Human-Centric Video Generation via Collaborative Multi-Modal Conditioning. *arXiv preprint arXiv:2509.08519*.

Chen, M.; Chen, J.; Fan, Z.; Lee, Y.; Dang, Z.; Wang, L.; Cui, Y.; Chau, L.-P.; and Wang, Y. 2026. HVG-3D: Bridging Real and Simulation Domains for 3D-Conditional Hand-Object Interaction Video Synthesis. *arXiv preprint arXiv:2604.03305*.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Fan, Y.; Yang, Q.; Wang, K.; Zhou, H.; Li, Y.; Feng, H.; Ding, E.; Wu, Y.; and Wang, J. 2025. Re-HOLD: Video Hand Object Interaction Reenactment via Adaptive Layout-Instructed Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17550–17560.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hu, T.; Yu, Z.; Zhou, Z.; Liang, S.; Zhou, Y.; Lin, Q.; and Lu, Q. 2025. HunyuanCustom: A Multimodal-Driven Architecture for Customized Video Generation. *arXiv preprint arXiv:2505.04512*.

Huang, X.; Xiang, M.; Shen, Z.; Wu, J.; Wu, C.; Zhao, C.; Wang, K.; Zhou, H.; Liu, S.; Feng, H.; et al. 2026. GenHOI: Towards Object-Consistent Hand-Object Interaction with Temporally Balanced and Spatially Selective Object Injection. *arXiv preprint arXiv:2603.06048*.

Huang, Z.; Zhou, Z.; Cao, J.; Ma, Y.; Chen, Y.; Rao, Z.; Xu, Z.; Wang, H.; Lin, Q.; Zhou, Y.; et al. 2025. HOMA: Towards Generic Human-Object Interaction in Multimodal Driven Human Animation with Weak Conditions. In *SIGGRAPH Asia 2025 Conference Papers*, 1–12.

Jiang, T.; Xie, X.; and Li, Y. 2024. RTMW: Real-time multi-person 2D and 3D whole-body pose estimation. *arXiv preprint arXiv:2407.08634*.

Jiang, Z.; Han, Z.; Mao, C.; Zhang, J.; Pan, Y.; and Liu, Y. 2025. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17191–17202.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. HunyuanVideo: A Systematic Framework for Large Video Generative Models. *arXiv preprint arXiv:2412.03603*.

Lin, H.; Chen, S.; Liew, J. H.; Chen, D. Y.; Li, Z.; Zhao, Y.; Peng, S.; Guo, H.; Zhou, X.; Shi, G.; Feng, J.; and Kang, B. 2026. Depth Anything 3: Recovering the Visual Space from Any Views. In *International Conference on Learning Representations*.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.

Liu, L.; Ma, T.; Li, B.; Chen, Z.; Liu, J.; Li, G.; Zhou, S.; He, Q.; and Wu, X. 2025. Phantom: Subject-Consistent Video Generation via Cross-Modal Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14951–14961.

Liu, X.; Gong, C.; and Liu, Q. 2023. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*.

Pang, Y.; Shao, R.; Zhang, J.; Tu, H.; Liu, Y.; Zhou, B.; Zhang, H.; and Liu, Y. 2025. ManiVideo: Generating Hand-Object Manipulation Video with Dexterous and Generalizable Grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12209–12219.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2025. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*.

Ressler-Antal, T.; Fundel, F.; Alaya, M. B.; Baumann, S. A.; Krause, F.; Gui, M.; and Ommer, B. 2025. DisMo: Disentangled Motion Representations for Open-World Motion Transfer. In *Advances in Neural Information Processing Systems*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, Z.; Wu, C.; Zhou, J.; Zhao, C.; Wang, K.; Zhou, H.; Li, Y.; Feng, H.; He, W.; and Wang, J. 2025. iDiT-HOI: Inpainting-based Hand Object Interaction Reenact-

ment via Video Diffusion Transformer. *arXiv preprint arXiv:2506.12847*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025a. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, L.; Xia, Z.; Hu, T.; Wang, P.; Wei, P.; Zheng, Z.; Zhou, M.; Zhang, Y.; and Gao, M. 2025b. Dreamactor-h1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers. *arXiv preprint arXiv:2506.10568*.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

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

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7623–7633.

Xu, Z.; Huang, Z.; Cao, J.; Zhang, Y.; Cun, X.; Shuai, Q.; Wang, Y.; Bao, L.; and Tang, F. 2026a. AnchorCrafter: Animate Cyber-Anchors Selling Your Products via Human-Object Interacting Video Generation. *IEEE Transactions on Visualization and Computer Graphics*, 32(7): 5098–5112.

Xu, Z.; Rao, Z.; Cao, J.; Liu, X.; Fang, Z.; Zhang, H.; Tang, S.; and Tang, F. 2026b. GeoHOI: Geometry-Enhanced Human-Object Interaction Video Generation via Hierarchical Multi-Modal Injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3739–3748.

Xue, Z.; Luo, M.; Chen, C.; and Grauman, K. 2024. HOI-Swap: Swapping Objects in Videos with Hand-Object Interaction Awareness. *NeurIPS*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2025. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. In *International Conference on Learning Representations*.

Yu, T.; Wang, Z.; Wang, C.; Huang, F.; Ma, W.; He, Z.; Cai, T.; Chen, W.; Huang, Y.; Zhao, Y.; et al. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

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

Zhang, Y.; Gu, J.; Wang, L.-W.; Wang, H.; Cheng, J.; Zhu, Y.; and Zou, F. 2025a. MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. In *International Conference on Machine Learning*.

Zhang, Z.; Teng, J.; Yang, Z.; Cao, T.; Wang, C.; Gu, X.; Tang, J.; Guo, D.; and Wang, M. 2025b. Kaleido: Open-Sourced Multi-Subject Reference Video Generation Model. *arXiv preprint arXiv:2510.18573*.

A Implementation Details

A.1 Data Collection and Preprocessing

We collect approximately 100 hours of real-world HOI videos, organized as 75K five-second clips and covering around 600 distinct products. Each video is associated with multiple product images depicting the interacted object from different viewpoints. We first use SAM2 (Ravi et al. 2025) to segment the interacted object throughout the video and extract a sequence of object crops. RTMW (Jiang, Xie, and Li 2024) is used to obtain human-pose conditions, and a vision-language model (Yu et al. 2025) generates the corresponding video descriptions.

For the synthetic cross-object data, we construct training pairs from 40,000 Objaverse objects (Deitke et al. 2023). In each pair, the source and target correspond to different object instances. For each object, we render 20 motion frames; the source sequence provides the driving motion, while the target object is rendered under the corresponding motion trajectory to provide supervision. We additionally render four reference images of the target object from distinct viewpoints. Consequently, the source motion and target appearance cannot be matched through identity copying: the model must transfer the motion encoded from the source sequence to the geometrically different target object by reasoning over its multi-view observations.

A.2 Network Architecture Details

Motion-condition feature adapter. The adapter is a lightweight convolutional modulation module. The image encoder $\mathcal{E}$ uses two 3×3 convolutional layers with ReLU activations to map the current three-channel object state to a 128-channel feature map. The DisMo motion extractor produces a 128-dimensional motion embedding, which the linear modulation layer $\mathcal{H}$ maps to channel-wise scale and shift parameters. These parameters modulate the RMS-normalized image feature as defined in the main paper. The decoder $\mathcal{D}$ mirrors the encoder with two 3×3 convolutional layers and maps the modulated feature back to the three-channel motion query P_t. The motion extractor, $\mathcal{E}$, $\mathcal{H}$, and $\mathcal{D}$ are jointly optimized with MDOP.

RGB prediction head. Following the dense-prediction design of VGGT (Wang et al. 2025a) and DA3-Large (Lin et al. 2026), the RGB head uses the same four intermediate backbone layers as DA3-Large, indexed by $\{11, 15, 19, 23\}$. We retain the tokens associated with the motion-query view and decode them through the standard DPT multi-scale projection and coarse-to-fine residual fusion pathway. The fused feature is bilinearly upsampled to the input resolution and mapped to four output channels with a lightweight convolutional decoder. After a sigmoid activation, the first three channels form the predicted RGB anchor $\hat{O}_{t+\Delta}$, while the fourth is an auxiliary confidence channel; only the RGB prediction is used in the autoregressive anchor rollout.

A.3 Training Details

Stage-I training. All object crops and target reference images are processed at a resolution of 266×266. The implicit motion encoder is initialized from DisMo (Ressler-Antal et al. 2025). The multi-view fusion backbone of the 3D foundation model is initialized from DA3-Large (Lin et al. 2026), while the motion-condition feature adapter, the adapted multi-view backbone, and the RGB prediction head are jointly optimized. We train these modules using a mixture of real-HOI self-reconstruction samples and synthetic Objaverse cross-object samples. Training is conducted on 8 NVIDIA A800 (80GB) GPUs for 10,000 iterations with a batch size of 64 and a fixed learning rate of 6×10^{-4}. The reconstruction-loss weights are set to $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$, respectively.

Stage-II training. The video-generation backbone is initialized from Wan2.1-T2V-14B (Wan et al. 2025). Following VACE (Jiang et al. 2025), we initialize the additional context branch by cloning the corresponding pretrained backbone blocks and train it on the collected real HOI dataset using the proxy-guidance construction described in the main paper. Stage II is trained on 81-frame clips at 720p resolution for 20,000 iterations on 16 NVIDIA A800 (80GB) GPUs with a per-GPU batch size of 1 (effective batch size 16), a learning rate of 1×10^{-5}, and bf16 mixed precision.

Inference details. Videos are generated with 50 sampling steps and a classifier-free guidance scale of 5.0. The object mask is expanded before conditioning, leaving room for target objects whose shape differs from the source. The attention enhancement operates only at inference and introduces no additional training cost. Long videos are generated in segments of up to 81 frames; quantitative and qualitative results are provided in Sec. C.4.

B Evaluation Protocol

B.1 Evaluation Datasets

We evaluate MVHOI on three complementary test sets covering synthetic cross-object transfer, real-world self-reconstruction, and real-world cross-object reenactment.

First, we construct a held-out synthetic evaluation set from Objaverse (Deitke et al. 2023) to evaluate MDOP independently. The source and target objects in each pair have different identities, and none of the selected objects overlap with the Stage-I training set. Because the target object can be rendered under the same motion trajectory as the source object, this test set provides frame-aligned ground truth for evaluating cross-object motion transfer.

Second, we select 100 held-out HOI videos from our collected real-world dataset. Each video is accompanied by multi-view images of the product appearing in the interaction. These videos and products are excluded from the data used to train both stages. We use this subset for Self-Reenactment by conditioning the model on multi-view images of the same product and evaluating its reconstruction of the original HOI video.

Third, we evaluate real-world Cross-Reenactment using source videos paired with multi-view images of different target products. This test set contains 142 fixed source–target pairs: 29 representative source sequences from the publicly released AnchorCrafter dataset (Xu et al. 2026a) and 113

pairs constructed from our held-out HOI data. The source and target identities are always different in this setting. Since AnchorCrafter provides three product views, all methods use the same three reference images on this subset. For our collected data, four reference views are used when available. The source–target pairs are fixed in advance and shared by all evaluated methods.

B.2 Evaluation Metrics

Generated and ground-truth videos are matched by sample identifier and decoded in RGB. For frame-aligned metrics, we resize the ground-truth frames to the spatial resolution of the generated video, using area interpolation for downsampling and bicubic interpolation for upsampling. If the two videos have different lengths, the ground-truth sequence is truncated to the generated length or padded by repeating its final frame. Unless otherwise stated, the reported score is the mean over all evaluated frames and videos.

Frame-aligned reconstruction metrics. We compute PSNR (Al Najjar 2024) on full RGB frames in the $[0, 255]$ range with a peak value of 255. SSIM (Wang et al. 2004) is computed on RGB frames with a data range of 255 using the channel-aware implementation from `scikit-image`. LPIPS (Zhang et al. 2018) uses the learned AlexNet backbone; RGB inputs are normalized from $[0, 255]$ to $[-1, 1]$ before feature extraction. These three metrics are used only when frame-aligned ground truth is available.

Distributional video-quality metrics. FID (Heusel et al. 2017) compares the pooled distributions of all generated and corresponding real-video frames. We use the 2048-dimensional Inception-V3 representation implemented by `torchmetrics`, process RGB frames as 8-bit tensors, and compute the Fréchet distance from the empirical means and covariance matrices of the two frame pools. KID (Bińkowski et al. 2018), used for the Stage-I synthetic Cross-Object Transfer experiment, is computed from the same Inception representation with subsets of 50 samples, and we report the estimated mean kernel distance.

For FVD (Unterthiner et al. 2018), each video is decoded in its entirety, bicubically resized to 224×224, and normalized to $[-1, 1]$. We extract one video-level representation with an I3D network pretrained on Kinetics-400 and compute the Fréchet distance between the generated- and real-video feature distributions. Generated and real videos are paired using the same fixed evaluation list, and all compared methods are evaluated with the same temporal extent.

Object fidelity. O-CLIP (Xu et al. 2026a; Huang et al. 2025) is computed with the OpenCLIP ViT-B/16 image encoder pretrained on LAION-2B (`laion2b_s34b_b88k`). For every generated frame, we use the corresponding ground-truth object mask, binarized at 127, to remove background pixels and derive a tight object bounding box. The crop is zero-padded to a square and bilinearly resized to 224×224. Both the crop and each available target-object reference view are encoded into ℓ_2-normalized CLIP features. We compute their cosine similarities, select the maximum across reference views for each frame, and average these frame-level maxima over each video and then over the evaluation set. This maximum-over-views aggregation allows the metric to match each generated object state to its most compatible observed target view.

B.3 Human Study Protocol

In addition to automatic metrics, we conduct a human study along three axes: **Motion Consistency (MC)**, **Visual Quality (VQ)**, and **Human–Object Interaction Realism (HR)**. MC measures whether the synthesized object follows the driving motion and remains synchronized with the hand throughout manipulation (e.g., avoiding drift, slipping, or temporal jitter). VQ assesses overall perceptual quality, including sharpness, visible artifacts, and temporal smoothness. HR evaluates the physical plausibility and naturalness of the interaction, with particular emphasis on contact correctness and realistic hand–object coupling under occlusions and large viewpoint changes. The study involves 30 participants and covers all 142 short-video Cross-Reenactment pairs and 113 long-video test cases. Videos are presented in randomized order with method identities concealed. All 30 participants rate every generated video from each method included in the user study along all three axes on a 1–5 scale (5 indicates the best quality), and we report the mean score over all participants and test cases.

B.4 Baseline Adaptation Details

We adapt inputs according to each baseline's native conditioning interface, as specified below. All methods generate videos at matched spatial resolution, temporal length, and frame rate. Unless noted otherwise, we use official checkpoints and recommended inference settings. No test sequence or target product is used to train or fine-tune any evaluated model.

MimicMotion. MimicMotion is a pose-guided human animation method and does not natively support object replacement. To prepare its single-image input, we use Google's Nano Banana image-editing model (November 2025 version) in an offline preprocessing step. Given the first frame of the source interaction video and a target-object reference image, it edits the first frame to contain the target product. We then provide the edited first frame and the human-pose sequence extracted from the source interaction video to MimicMotion. Nano Banana is used only to prepare this initial condition and is not invoked during MimicMotion video generation; all subsequent frames are generated through MimicMotion's original pose-control pathway.

VACE. VACE is the closest controllable-editing baseline to our Stage-II setup. We feed it the same masked source video and available structural conditions used by our video-generation stage (e.g., depth and related structure maps when supported). Because the official checkpoint is not specialized for multi-view HOI reenactment with product references, we additionally fine-tune VACE on our collected multi-view HOI dataset with multiple reference views, under the same data split used for MVHOI. This fine-tuning is applied only to VACE among the compared methods, so that the comparison

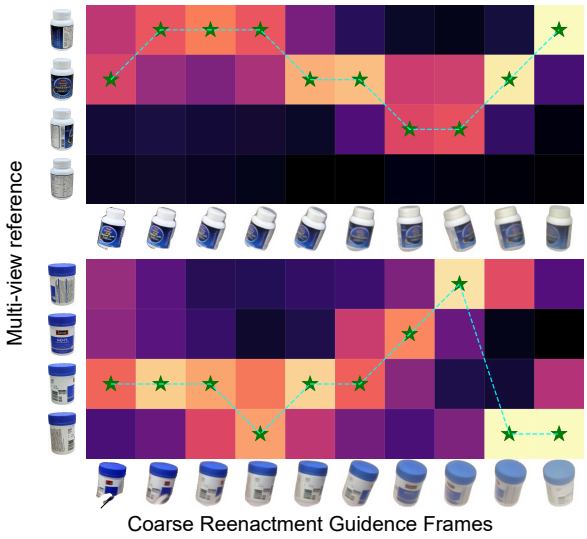

Figure 5: Cross-view association weights between motion-conditioned query states and target-object reference views.

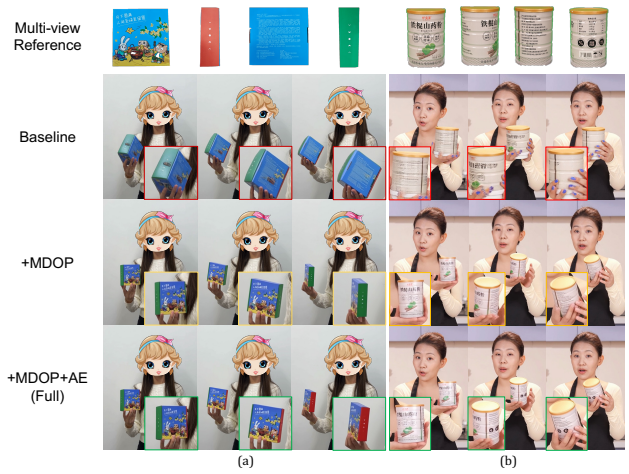

Figure 6: Qualitative comparison of the baseline, MDOP conditioning, and the full model with attention enhancement (AE).

reflects a strong domain-adapted controllable baseline rather than an out-of-domain official checkpoint.

HunyuanCustom. HunyuanCustom is evaluated as a multi-subject reference-conditioned generator. We provide the masked source video together with a front-view target-product image as the object reference, following its recommended subject-conditioning protocol.

HuMo. HuMo receives a human reference image, target-product reference images, and the corresponding text description. This matches its native multi-subject conditioning interface and allows the model to condition on both the interacting person and the replacement object.

GenHOI. GenHOI is an HOI-oriented reenactment baseline. We follow its official inference protocol and provide the source interaction video together with target-product references, without modifying its training setup.

C Additional Experiment Results

C.1 Cross-View Association Visualization

Appearance-driven attention in a video diffusion model may favor a visually similar but geometrically incompatible reference, particularly during large rotations or hand–object occlusions. To examine the view-selection signal formed inside MDOP, Fig. 5 visualizes the normalized association weights between each motion-conditioned query state and the target-object reference views. Each row corresponds to a query state along the object trajectory, while the columns correspond to the available reference viewpoints.

As the visible side of the object changes, the attention mass shifts toward geometrically compatible reference views rather than remaining fixed on a single appearance reference. Intermediate object states also distribute weight across adjacent views instead of making an abrupt hard selection. This visualization suggests that MDOP captures a correspondence between the evolving object state and the reference viewpoints. We reuse these associations as a soft attention bias in Stage II, where they complement appearance similarity and reduce reference-view confusion.

C.2 Qualitative Ablation Results

Figure 6 complements the quantitative ablation in the main paper with two representative interactions. Within each example, the source interaction and multi-view target references are held fixed across the three rows, isolating the effect of the added conditions. The baseline conditions the video generator only on the multi-view target references. Although these references provide object appearance, they do not specify the frame-dependent object state; consequently, the generated object may follow an incorrect orientation, change shape across frames, or become misaligned with the driving interaction.

In the shown examples, adding the coarse anchors produced by MDOP supplies an explicit frame-dependent structural condition. The middle row follows the source manipulation more closely and preserves a more stable silhouette and object scale under rotation. However, the coarse anchors do not retain all high-frequency appearance details, so the video generator can still confuse reference viewpoints.

The full model additionally applies AE to favor references compatible with the current object state. As shown in the bottom row, it better preserves view-dependent colors, labels, and surface details while retaining the motion and geometry established by MDOP. Together, these qualitative results distinguish the two components: MDOP principally establishes the coarse motion-aligned object state, whereas AE refines the selection of appearance information from the target views. The progression is consistent with the quantitative ablation in the main paper.

Method	FID ↓	FVD ↓	O-CLIP ↑
MimicMotion	87.25	892.2	0.421
VACE	79.23	618.2	0.592
HuMo	277.20	2530.0	0.333
HunyuanCustom	84.56	699.0	0.608
GenHOI	72.15	578.2	0.613
Ours	**66.44**	**534.1**	**0.617**

Table 5: Automatic-metric comparison on the AnchorCrafter benchmark. FID/FVD measure perceptual realism and temporal coherence (lower is better), while O-CLIP evaluates object appearance consistency (higher is better).

Figure 7: Qualitative comparison between vanilla autoregressive (AR) inference and cross-iterative inference, shown at an interval of 80 frames. Frames with human context in the bottom row denote refined outputs used at segment boundaries.

C.3 Results on the AnchorCrafter Benchmark

As reported in Tab. 5, MVHOI obtains the best value on each of the three automatic metrics. The consistent gains in distributional video quality, temporal coherence, and target-object fidelity indicate that the benefits of motion-aligned structural guidance and multi-view appearance conditioning extend to the public AnchorCrafter benchmark rather than only to the collected data.

C.4 Long-Video Generation

Comparison with baselines. We evaluate the complete framework on 10-second Cross-Reenactment sequences using matched source videos and target-product conditions. As reported in Tab. 6, MVHOI achieves the best performance across all three automatic metrics. The consistent gains indicate that MVHOI retains both target-object appearance and video-level coherence as the generation horizon increases.

This comparison evaluates the complete MVHOI system rather than isolating the contribution of cross-iterative inference. In particular, each segment is generated from MDOP anchors and multi-view references, and the refined object state at one segment boundary initializes the next MDOP rollout. The result therefore reflects the combined effect of motion-aware prior construction, reference conditioning, and cross-iterative state refresh.

Effect of cross-iterative inference. As illustrated in Figure 7, vanilla AR gradually exhibits geometric distortion and severe object deformation, whereas cross-iterative inference better preserves the object's shape and texture over the same horizon. The refined frames used at segment boundaries provide subsequent MDOP rollouts with a higher-quality visual state than a purely autoregressive continuation, reducing the accumulation of coarse-anchor errors over successive clips.

Method	FID ↓	FVD ↓	O-CLIP ↑
VACE	52.47	279.1	0.546
HunyuanCustom	61.48	364.4	0.532
GenHOI	48.08	281.6	0.549
Ours	**38.15**	**267.0**	**0.557**

Table 6: Quantitative comparison with baseline methods on long-video generation. The corresponding human evaluation is reported in the user-study table of the main paper.

D Limitations

Viewpoint coverage. The synthesis quality depends on the comprehensiveness of the multi-view references. If an object rotates to a viewpoint not covered by the reference set, the model may produce blurred textures or inconsistent appearances due to insufficient visual cues.

Motion extraction under occlusion. Extreme or prolonged hand–object occlusions in the source video can hinder the motion extractor. This may lead to inaccurate 3D trajectory distillation, resulting in misaligned poses or physically implausible interactions.

Dependence on object masks. The current pipeline requires object masks to extract source-object crops and to condition the masked video-generation process. Inaccurate segmentation or tracking can contaminate the motion signal, leave residual source-object content, or exclude regions needed for the replacement object. Integrating robust automatic mask estimation and uncertainty-aware conditioning would make the system more practical for unconstrained videos.