EXO2EGOHOI: HAND-OBJECT-INTERACTION AWARE EXOCENTRIC-TO-EGOCENTRIC VIDEO GENERATION

Anonymous authors
Paper under double-blind review

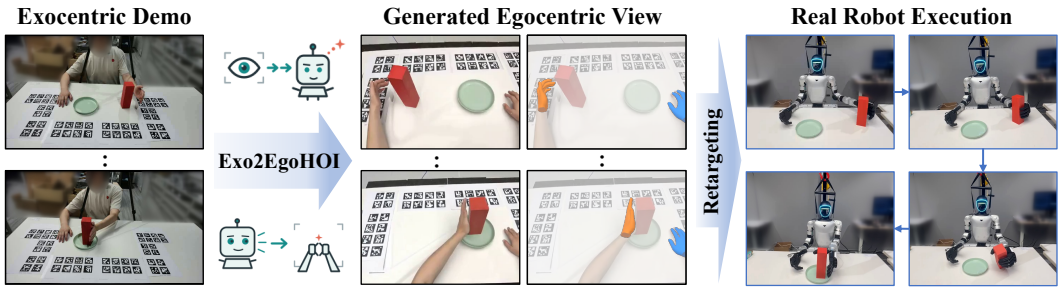


Figure 1: **From exocentric human demonstrations to egocentric observations and robot execution.** Exo2EgoHOI preserves hand-object interactions during exocentric-to-egocentric video generation, enabling downstream robot execution via retargeting.

ABSTRACT

Egocentric videos of human manipulation provide valuable visual experience for embodied intelligence, yet collecting such data at scale is costly. Exocentric-to-egocentric video generation offers a scalable alternative by transforming abundant third-person manipulation videos into first-person observations. However, existing methods often struggle to faithfully preserve demonstrated hand-object interactions (HOI) across large viewpoint changes due to insufficient fine-grained interaction guidance and weak object-centric anchoring. We present Exo2EgoHOI, an HOI-aware video generative framework for interaction-preserving exocentric-to-egocentric translation. To preserve fine-grained HOI, we introduce a unified 4D HOI prior that combines scene geometry, articulated hand renderings, and dense hand-object relation fields, together with a dual-branch residual adapter for injecting structural and relational cues into the video generation backbone. To preserve object consistency, we introduce Decomposed Gated Cross-Attention, which separately encodes object and background references and adaptively integrates global semantic and local appearance features as object-centric anchors. Experiments on ARCTIC-HOI and Ego-Exo4D demonstrate substantial improvements in object consistency and HOI preservation while maintaining competitive visual fidelity. In particular, on ARCTIC-HOI, Exo2EgoHOI improves object mIoU by 32.3% and reduces MPJPE and PA-MPJPE by 34.7% and 50.0%, respectively, relative to the respective best baseline results. Code will be released to benefit the research community.

1 INTRODUCTION

Egocentric videos of human manipulation provide valuable visual experience for embodied intelligence (Wang et al., 2026c; Zhou et al., 2026; Zheng et al., 2026; Yang et al., 2025; Lu et al., 2026), capturing hand movements, object changes, and actions from the actor's perspective. Such data are commonly collected using wearable cameras (Grauman et al., 2022; Kareer et al., 2025), but scaling collection across diverse people, environments, and tasks requires substantial effort. In contrast, the Internet contains abundant third-person footage of everyday human manipulation. This motivates *exocentric (source)-to-egocentric (target) video translation*, which converts readily available third-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

person videos into first-person visual experiences. As illustrated in Fig. 1, the generated egocentric observations can further provide an interface for transferring human manipulation to robotic systems. Building on recent advances in video generation and view-controllable synthesis (Lin et al., 2026b; Yu et al., 2025; Wan et al., 2025), recent methods have begun exploring such viewpoint translation (Liu et al., 2024a; Mahdi et al., 2025; Kang et al., 2026; Park et al., 2026; Wang et al., 2026a).

Despite recent progress, faithfully translating manipulation videos remains challenging: existing methods may produce visually plausible target-view videos while altering the demonstrated *hand-object interactions (HOI)*, such as finger configurations or hand-object contacts. We identify two key challenges. First, *scene-level geometric conditioning is insufficient to preserve fine-grained interaction geometry*. Hands and manipulated objects occupy only a small, often occluded portion of the scene, leaving fine-grained finger configurations and local hand-object relationships poorly resolved despite scene-level geometric constraints. Second, *maintaining manipulated-object consistency is challenging under large exocentric-to-egocentric viewpoint changes*. Such viewpoint changes expose object regions unseen in the source view, making target-view appearance only partially constrained by geometry-based warping and rendering (Yu et al., 2025). Meanwhile, global image conditioning lacks explicit object-centric correspondence, diluting object-specific cues with dominant background information. As a result, the manipulated object may drift in appearance, shape, or state throughout generation. Faithful HOI-preserving translation therefore requires both fine-grained interaction and persistent object-centric anchoring throughout cross-view video generation.

To address these challenges, we introduce Exo2EgoHOI, an HOI-aware video generative framework for exocentric-to-egocentric viewpoint translation (Fig. 2). First, we introduce a *unified 4D HOI prior* that augments scene-level geometry with temporally aligned articulated hand renderings and dense hand-object relation fields in the target view. The hand renderings specify target-view hand poses and finger configurations, while the relation fields encode local proximity and relative direction between the hands and the manipulated object. A dual-branch residual HOI adapter separately encodes these structural and relational cues and fuses them into residual features, providing explicit fine-grained interaction guidance to the video generative model. Second, we introduce *Decomposed Gated Cross-Attention* to preserve object consistency across viewpoints and throughout video generation. Instead of conditioning on the source image as a single global reference, we separately encode the manipulated object and background from the first exocentric frame, providing the object with a dedicated source-view anchor. Their global semantic and local appearance features are adaptively integrated through learned per-block fusion, preserving object-specific characteristics while retaining complementary scene context. To systematically evaluate HOI-aware viewpoint translation, we further construct an *interaction-oriented evaluation benchmark* based on ARCTIC (Fan et al., 2023) and a collection of in-the-wild HOI videos. The benchmark jointly evaluates visual fidelity, object consistency, and hand-pose consistency, assessing both overall generation quality and preservation of the demonstrated hand-object interactions.

In summary, our main contributions are: (i) We propose Exo2EgoHOI, an HOI-aware exocentric-to-egocentric video generation framework, featuring a *unified 4D HOI prior* and dual-branch residual adapter for preserving fine-grained hand-object interactions under large viewpoint changes. (ii) We introduce *Decomposed Gated Cross-Attention*, which separately models object and background references and adaptively integrates global semantic and local appearance cues to preserve manipulated-object consistency across viewpoints and throughout video generation. (iii) Extensive experiments on ARCTIC-HOI and Ego-Exo4D demonstrate substantial improvements in object consistency and interaction preservation while maintaining strong visual fidelity, with comprehensive ablations validating the proposed components.

2 Related Work

Exocentric-to-Egocentric Video Generation. Recent controllable video generation methods incorporate camera trajectories and rendered scene geometry into pretrained diffusion models to enable viewpoint and camera-motion control (He et al., 2025; Bahmani et al., 2025; Lin et al., 2026b; Yu et al., 2025; Team et al., 2026). Together with large-scale egocentric and synchronized ego-exo datasets (Grauman et al., 2022; 2024), these advances have facilitated exocentric-to-egocentric generation. Exo2Ego-V (Liu et al., 2024a) and Exo2EgoSyn (Mahdi et al., 2025) exploit synchronized external views, while recent methods increasingly address single-view inputs. EgoX (Kang et al.,

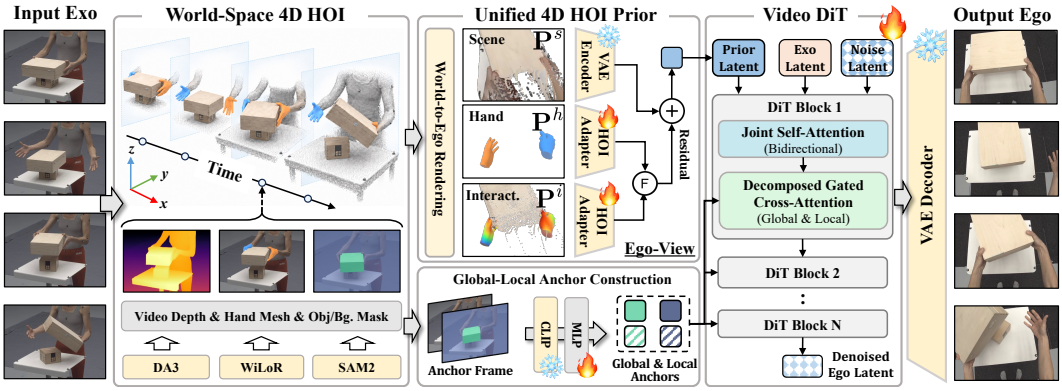


Figure 2: **Overview of Exo2EgoHOI.** Given an exocentric video, we reconstruct world-space 4D HOI and transform it into a unified egocentric prior encoding scene geometry, hand motion, and interaction cues. HOI adapters inject this prior into a video Diffusion Transformer (DiT), while Decomposed Gated Cross-Attention anchors object-specific semantic and appearance cues from the exocentric reference to preserve object consistency throughout egocentric video generation.

2026) uses rendered scene priors and geometry-guided attention, EgoWorld (Park et al., 2026) incorporates geometry and hand-pose conditioning, and Grounded-Exo2Ego (Wang et al., 2026a) introduces object-level semantic grounding to complement incomplete geometric evidence. Despite these advances, existing conditioning primarily focuses on scene-level alignment or coarse object grounding, while fine-grained interaction preservation and persistent object-centric anchoring remain insufficiently addressed. Our work instead targets *interaction-preserving* viewpoint translation by explicitly modeling HOI structure throughout the video.

Hand-Object Interaction Modeling and Generation. Accurate HOI modeling requires not only articulated hand geometry but also the spatial and contact relationships between hands and manipulated objects. Parametric hand models and recent reconstruction methods (Romero et al., 2017; Pavlakos et al., 2024; Zeng et al., 2026b; Potamias et al., 2025) provide strong priors for recovering hand articulation, while datasets such as ARCTIC (Fan et al., 2023) support detailed 3D modeling of interacting hands and objects. Beyond hand pose alone, recent approaches explicitly model hand-object relations for interaction reconstruction and synthesis. GRIP (Taheri et al., 2024) exploits spatiotemporal body-object relations, GeneOH Diffusion (Liu & Yi, 2024) introduces contact-centric representations of hand-object geometry, and DiffH2O (Christen et al., 2024) jointly models hand and object motion for interaction synthesis. More recent generative approaches (Zeng et al., 2026a; Li et al., 2026; Wang et al., 2026b) further incorporate hand kinematics or contact structure into dynamic HOI generation. These works demonstrate the importance of explicit interaction cues beyond visual appearance or hand pose alone. In contrast, our goal is not to synthesize a new plausible interaction, but to *preserve an observed interaction across viewpoints*.

3 Method

3.1 Problem Formulation

As shown in Fig. 2, given an exocentric hand-object interaction video $\mathbf{X}^x = \{\mathbf{I}_t^x\}_{t=1}^T$, our goal is to generate its corresponding egocentric sequence $\mathbf{X}^e = \{\mathbf{I}_t^e\}_{t=1}^T$ while faithfully preserving the observed HOI. We condition the generation on a text description l, camera calibration $\mathcal{C}$, a unified ego-view *4D HOI prior* $\mathbf{P}^{4D}$, and the object and background masks $(\mathbf{M}_1^o, \mathbf{M}_1^b)$ from the first exocentric frame. The conditional generation problem is formulated as:

$$p_\theta\big(\mathbf{X}^e \mid \mathbf{X}^x, l, \mathcal{C}, \mathbf{P}^{4D}, \mathbf{M}_1^o, \mathbf{M}_1^b\big)\,, \tag{1}$$

where $\mathbf{P}^{4D} = \{\mathbf{P}^s, \mathbf{P}^h, \mathbf{P}^i\}$ consists of the scene, hand, and interaction priors, respectively. $\mathcal{C}$ contains the camera intrinsics and extrinsics for the exocentric view, $\mathbf{K}^x$ and $\{\mathbf{T}_{wc,t}^x\}$, and those for the target egocentric view, $\mathbf{K}^e$ and $\{\mathbf{T}_{wc,t}^e\}$.

3.2 UNIFIED 4D HOI PRIOR

Scene-level geometry provides useful spatial constraints for viewpoint translation, but does not explicitly characterize fine-grained hand-object interactions. We therefore construct a unified 4D HOI prior with three complementary components: a scene prior $\mathbf{P}^s$ describing the global scene layout, a hand prior $\mathbf{P}^h$ specifying articulated hand geometry, and an interaction prior $\mathbf{P}^i$ encoding local hand-object spatial relations. Together, they provide progressively finer constraints on scene structure, hand articulation, and hand-object interaction throughout the target egocentric sequence.

World-Space 4D HOI Reconstruction. Given an exocentric HOI video $\mathbf{X}^x$, we first reconstruct the observed scene, hands, and manipulated object in a shared world coordinate system using off-the-shelf models. Specifically, DA3 (Lin et al., 2026a) estimates temporally consistent depth maps, WiLoR (Potamias et al., 2025) estimates MANO (Romero et al., 2017) hand parameters, and SAM2 (Ravi et al., 2025) provides object and background masks:

$$D_t^x = \mathrm{DA3}(\mathbf{X}^x)_t, \quad (\boldsymbol{\beta}_t, \boldsymbol{\theta}_t, \boldsymbol{\tau}_t) = \mathrm{WiLoR}(\mathbf{X}^x)_t, \quad (\mathbf{M}_t^o, \mathbf{M}_t^b) = \mathrm{SAM2}(\mathbf{X}^x)_t. \tag{2}$$

Here, t indexes frames, and $\boldsymbol{\beta}_t$, $\boldsymbol{\theta}_t$, and $\boldsymbol{\tau}_t$ denote hand shape, pose, and translation, respectively. We back-project the depth maps into scene point clouds $\mathcal{P}_t^s$ and extract object point clouds $\mathcal{P}_t^o$ using the object masks $\mathbf{M}_t^o$. Given a pixel $\mathbf{u} = [u, v]$ in the exocentric view and its depth $D_t^x(u, v)$, the corresponding scene point is transformed into the shared world frame using the exocentric camera intrinsics and extrinsics:

$$\mathcal{P}_t^s(u, v) = \mathbf{T}_{wc,t}^x [D_t^x(u, v)(\mathbf{K}^x)^{-1}\mathbf{u}; 1]. \tag{3}$$

The resulting temporal sequence forms a *world-space 4D HOI reconstruction*, from which scene, hand, and interaction cues can be consistently projected into the egocentric view using $\mathbf{K}^e$, $\mathbf{T}_{wc,t}^e$.

Ego-View Scene Prior. Similar to (Yu et al., 2025), we render the scene point cloud $\mathcal{P}_t^s$ from the target egocentric camera at each time step and stack the rendered frames to construct the scene priors:

$$\mathbf{P}^s = \left\{\mathcal{R}_s(\mathcal{P}_t^s; \mathbf{K}^e, \mathbf{T}_{wc,t}^e)\right\}_{t=1}^T \in \mathbb{R}^{3\times T\times H\times W}, \tag{4}$$

where $\mathcal{R}_s$ denotes point-cloud rendering (Ravi et al., 2020). The scene prior provides the global layout and coarse visual context in the target view. However, because hands and manipulated objects are small and frequently occluded, scene geometry alone provides only incomplete constraints on fine-grained HOI. We therefore complement it with explicit hand and interaction priors.

Ego-View Hand Prior. To explicitly constrain hand articulation, we construct the hand prior from the complete MANO hand mesh (Romero et al., 2017), including hand surfaces that may be occluded in the source view. Given the estimated hand shape $\boldsymbol{\beta}_t$, pose $\boldsymbol{\theta}_t$, and wrist translation $\boldsymbol{\tau}_t$, we obtain the world-space hand vertices $\mathcal{V}_t^h$ and mesh faces $\mathcal{F}_t^h$ as: $\mathcal{V}_t^h, \mathcal{F}_t^h = \mathbf{T}_{wc,t}^x \left[\mathrm{MANO}(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) + \boldsymbol{\tau}_t\right]$. We then render the hand meshes with assigned RGB colors $\mathbf{C}_t$ into the target egocentric view:

$$\mathbf{P}^h = \left\{\mathcal{R}_h\big(\mathcal{V}_t^h, \mathcal{F}_t^h, \mathbf{C}_t; \mathbf{K}^e, \mathbf{T}_{wc,t}^e\big)\right\}_{t=1}^T \in \mathbb{R}^{3\times T\times H\times W}, \tag{5}$$

where the hand renderer $\mathcal{R}_h$ uses a hand-only z-buffer and barycentric interpolation to produce per-pixel colors. The resulting prior explicitly specifies target-view hand shape and articulation beyond what can be recovered from the partial scene geometry.

Ego-View Interaction Prior. While $\mathbf{P}^h$ specifies hand articulation, it does not explicitly describe how the hand interacts with the manipulated object. We therefore characterize the local hand-object relation at each hand vertex using three complementary quantities: contact likelihood, hand-object distance, and relative direction. For each world-space hand vertex $\mathbf{v}_{t,i}^w \in \mathcal{V}_t^h$, we identify its five nearest object points in $\mathbf{p}_{t,j}^o \in \mathcal{P}_t^o$ and denote their index set by $\mathcal{N}_{t,i}$. The distance to each neighboring object point is $\delta_{t,i,j} = \|\mathbf{v}_{t,i}^w - \mathbf{p}_{t,j}^o\|_2$. We assign normalized inverse-distance weights and compute the weighted mean distance, where ϵ ensures numerical stability:

$$a_{t,i,j} = \max(\delta_{t,i,j}, \epsilon)^{-1} \Big/ \sum_{j'\in\mathcal{N}_{t,i}} \max(\delta_{t,i,j'}, \epsilon)^{-1}, \quad d_{t,i} = \sum_{j\in\mathcal{N}_{t,i}} a_{t,i,j}\delta_{t,i,j}. \tag{6}$$

Using the same weights, we compute the displacement from the hand vertex to the weighted center of its neighboring object points: $\mathbf{q}_{t,i}^w = \sum_{j\in\mathcal{N}_{t,i}} a_{t,i,j}\mathbf{p}_{t,j}^w - \mathbf{v}_{t,i}^w$. We then encode the local relation using a soft contact score $s_{t,i}$, normalized distance $\bar{d}_{t,i}$, and interaction direction $\mathbf{r}_{t,i}^w$:

$$s_{t,i} = \exp\left(-d_{t,i}^2/(2\sigma^2)\right), \quad \bar{d}_{t,i} = \mathrm{clip}(d_{t,i}/d_{\max}, 0, 1), \quad \mathbf{r}_{t,i}^w = \frac{\mathbf{q}_{t,i}^w}{\max(\|\mathbf{q}_{t,i}^w\|_2, \epsilon)}. \tag{7}$$

Here, σ controls the contact-score bandwidth, $d_{\max}$ sets the distance normalization range.

Using the rotation component $\mathbf{R}^e_{cw,t}$ of $\mathbf{T}^e_{cw,t}$, we transform the interaction direction into the target egocentric camera frame: $\mathbf{r}^e_{t,i} = \mathbf{R}^e_{cw,t}\mathbf{r}^w_{t,i}$. Each hand vertex is therefore associated with a five-dimensional descriptor $[s_{t,i}, \bar{d}_{t,i}, \mathbf{r}^e_{t,i}] \in \mathbb{R}^5$. Let $\mathcal{A}_t$ collect these descriptors over all hand vertices. We rasterize them into target egocentric view using the same hand topology as Eq. 5:

$$\mathbf{P}^i = \left\{\mathcal{R}_h\left(\mathcal{V}^h_t, \mathcal{F}^h_t, \mathcal{A}_t; \mathbf{K}^e, \mathbf{T}^e_{wc,t}\right)\right\}_{t=1}^{T} \in \mathbb{R}^{5\times T\times H\times W}. \tag{8}$$

The resulting $\mathbf{P}^i$ is pixel-aligned with $\mathbf{P}^s$ and $\mathbf{P}^h$, providing explicit target-view hand-object relational cues throughout the sequence.

Dual-Branch HOI Residual Adapter. We next integrate the complementary priors into the video generation backbone. The scene prior $\mathbf{P}^s$ is encoded by the frozen VAE of Wan (Wan et al., 2025), followed by patch embedding to obtain scene-conditioned egocentric tokens $\mathbf{Z}^e_{\text{in}}$. Since the RGB-like hand rendering $\mathbf{P}^h$ and the non-photometric relation field $\mathbf{P}^i$ have different signal characteristics, we encode them using two independent spatiotemporal branches, $\mathcal{F}_h$ and $\mathcal{F}_i$. Each branch consists of six 3D convolutional layers with SiLU activations, progressively downsampling the spatiotemporal resolution to match the video latent space. The two branches share the same architecture but use independent parameters, with input channels adapted to the 3-channel hand prior and 5-channel interaction prior, respectively. Detailed architecture is provided in the supplementary material.

The resulting feature grids are concatenated along the channel dimension, fused using a $1\times1\times1$ convolution $\mathcal{F}_{\text{fuse}}$, and projected by $\mathcal{P}$ to match the spatiotemporal resolution and channel dimension of $\mathbf{Z}^e_{\text{in}}$:

$$\Delta\mathbf{Z}^{\text{hoi}} = \mathcal{P}\left(\mathcal{F}_{\text{fuse}}\left([\mathcal{F}_h(\mathbf{P}^h); \mathcal{F}_i(\mathbf{P}^i)]\right)\right). \tag{9}$$

The projection $\mathcal{P}$ is zero-initialized such that the adapter initially produces a zero residual. We inject the fused HOI features into the scene-conditioned tokens through residual addition:

$$\mathbf{Z}^{4D} = \mathbf{Z}^e_{\text{in}} + \lambda_{\text{hoi}}\Delta\mathbf{Z}^{\text{hoi}}, \tag{10}$$

where λ_{hoi} is a learned residual scale initialized to one. The resulting HOI-conditioned tokens $\mathbf{Z}^{4D}$ integrate scene context, articulated hand structure, and explicit hand-object relations in a unified target-view representation.

3.3 Decomposed Gated Cross-Attention

Although the unified 4D HOI prior $\mathbf{P}^{4D}$ explicitly constrains interaction geometry, it does not guarantee that the manipulated object remains consistent across large viewpoint changes and throughout the generated sequence. Geometry-based rendering provides only partial target-view evidence, as previously unseen object surfaces cannot be recovered from the source observation. Moreover, conventional full-image conditioning can underrepresent the manipulated object, which typically occupies only a small region against a dominant background. We therefore introduce *Decomposed Gated Cross-Attention (DGCA)* to provide explicit object-centric anchoring during generation. DGCA separately encodes object and background references and adaptively integrates their global semantic and local appearance features through learned block-specific gates.

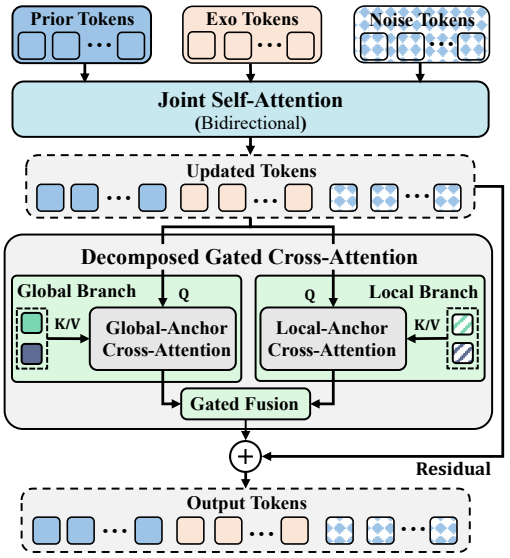


Figure 3: **Architecture of DiT block.** DGCA adaptively fuses global and local contexts through learnable block-specific gates.

Decomposed Global-Local Anchors. As illustrated in Fig. 2, given the first exocentric frame $\mathbf{I}^x_1$ and its object and background masks $(\mathbf{M}^o_1, \mathbf{M}^b_1)$ from Eq. 2, we extract an object crop

$\mathbf{I}_1^{x,o}$ and construct a background reference $\mathbf{I}_1^{x,b}$ with the manipulated object masked out. This regional decomposition provides the manipulated object with a dedicated representation while retaining complementary scene context.

We encode both references, $\mathbf{I}_1^{x,b}$ and $\mathbf{I}_1^{x,o}$, using a shared frozen CLIP encoder. For each region $r \in \{o, b\}$, the encoder produces a global `[CLS]` token $\mathbf{f}_{\text{cls}}^r$ and spatially resolved patch tokens $\mathbf{F}_{\text{patch}}^r$ via $[\mathbf{f}_{\text{cls}}^r; \mathbf{F}_{\text{patch}}^r] = \mathcal{E}_{\text{CLIP}}(\mathbf{I}_1^{x,r}), r \in \{o, b\}$. We then construct global and local reference anchors by concatenating the corresponding object and background features:

$$\mathbf{C}^g = [\mathbf{f}_{\text{cls}}^o; \mathbf{f}_{\text{cls}}^b], \qquad \mathbf{C}^\ell = [\mathbf{F}_{\text{patch}}^o; \mathbf{F}_{\text{patch}}^b], \tag{11}$$

The resulting representation is therefore decomposed along two complementary axes: *region*, separating the manipulated object from the background, and *feature granularity*, separating global semantics from local appearance. The global anchor $\mathbf{C}^g$ captures object and scene semantics, whereas the local anchor $\mathbf{C}^\ell$ retains spatially resolved appearance details. Together, they provide complementary cues for anchoring object identity and appearance throughout the video generation.

Gated Cross-Attention Fusion. As illustrated in Fig. 3, within each DiT block, bidirectional joint self-attention first processes exocentric video tokens, noisy egocentric tokens, and HOI-conditioned tokens. The resulting video features provide queries $\mathbf{Q}_i$ for reference cross-attention in the i-th block. The global and local anchors are passed through the shared image projector and the block-specific image key/value projections to obtain $(\mathbf{K}_i^r, \mathbf{V}_i^r)$ for $r \in \{g, \ell\}$. The corresponding cross-attention with d-dimensional attention heads is:

$$\mathbf{A}_i^r = \text{Attn}(\mathbf{Q}_i, \mathbf{K}_i^r, \mathbf{V}_i^r) = \text{softmax}\Big(\mathbf{Q}_i(\mathbf{K}_i^r)^\top/\sqrt{d}\Big)\,\mathbf{V}_i^r, \quad r \in \{g, \ell\}, \tag{12}$$

Rather than combining global and local appearance cues with fixed weights, we allow each DiT block to learn its own balance between the two feature granularities. Specifically, the i-th block learns a two-dimensional logit vector $\mathbf{a}_i$, which is normalized into gating coefficients:

$$\boldsymbol{\pi}_i = \text{softmax}(\mathbf{a}_i), \quad \mathbf{A}_i^{\text{img}} = \pi_i^g \mathbf{A}_i^g + \pi_i^\ell \mathbf{A}_i^\ell. \tag{13}$$

Here, π_i^g and π_i^ℓ control the contributions of global semantic and local appearance features, respectively. The gates are block-specific but shared across inputs and denoising steps, allowing different stages of the backbone to learn different global-local balances while the attention responses remain conditioned on the current video queries. The fused reference feature $\mathbf{A}_i^{\text{img}}$ is incorporated into the block's cross-attention update before the shared output projection and residual addition. In this way, DGCA complements the geometric 4D HOI prior with persistent object-centric anchoring, helping preserve object identity and appearance across viewpoints and throughout the generated sequence.

Training Objective. We train the model using conditional flow matching (Lipman et al., 2023). Given a paired egocentric training video $\mathbf{X}^e$, a frozen video VAE encodes it into a clean latent $\mathbf{z}_0^e = \mathcal{E}_{\text{VAE}}(\mathbf{X}^e)$. Let $\mathcal{K}$ collectively denote all conditioning inputs, including text prompt l, exocentric video context $\mathbf{X}^x$, unified 4D HOI prior $\mathbf{P}^{4D}$, and decomposed image reference anchors $\mathbf{C}^\ell$, $\mathbf{C}^g$. We sample a flow-matching timestep τ from the pretrained scheduler, obtain its corresponding noise level $\sigma_\tau \in [0, 1]$, and sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noisy latent is interpolated between the clean video latent and noise, and the model is trained to predict the corresponding latent velocity:

$$\mathbf{z}_\tau^e = (1 - \sigma_\tau)\mathbf{z}_0^e + \sigma_\tau \boldsymbol{\epsilon}, \qquad \mathcal{L}_{\text{FM}} = \mathbb{E}_{(\mathbf{X}^e, \mathcal{K}), \tau, \boldsymbol{\epsilon}}\left[\|\mathbf{v}_\theta(\mathbf{z}_\tau^e, \tau, \mathcal{K}) - (\boldsymbol{\epsilon} - \mathbf{z}_0^e)\|_2^2\right]. \tag{14}$$

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate on Ego-Exo4D (Grauman et al., 2024) using a setup similar to EgoX (Kang et al., 2026), covering diverse scenes and manipulation activities. Since Ego-Exo4D lacks a dedicated evaluation protocol for HOI-preserving video generation, we additionally construct ARCTIC-HOI, an interaction-oriented benchmark comprising 7K training clips and 180 non-overlapping test clips from ARCTIC (Fan et al., 2023), together with 60 in-the-wild HOI test clips.

Evaluation Metrics. We evaluate three aspects of the generated egocentric videos. For *visual fidelity*, we report PSNR, SSIM, LPIPS, and CLIP image similarity (CLIP-I). For *object consistency*, we

Table 1: **Quantitative comparison on ARCTIC-HOI and Ego-Exo4D datasets.** We evaluate visual fidelity, object consistency, and hand-pose consistency. Best results are shown in **bold** and second-best results are underlined.

	Visual Fidelity				Object Consistency			Hand Consistency		
Method	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	mIoU↑	Center-Err.↓	CLIP-O↑	MPJPE↓	WA-MPJPE↓	PA-MPJPE↓
ARCTIC-HOI										
WAN VACE (Jiang et al., 2025)	12.78	0.6163	0.5115	0.8485	0.2336	0.0893	0.8080	2870.22	371.18	20.36
EgoWorld (Park et al., 2026)	11.90	0.4652	0.6203	0.7995	0.2241	0.1131	0.7869	2282.03	440.60	57.92
Vista4D (Lin et al., 2026b)	**13.58**	0.6614	0.4324	0.8737	0.2768	0.0929	0.8176	2123.78	387.73	28.91
EgoX (Kang et al., 2026)	11.83	0.5941	0.4743	0.8910	0.2933	0.0804	0.8345	2019.81	325.56	22.12
Ours	13.43	**0.6685**	**0.3860**	**0.9175**	**0.3881**	**0.0583**	**0.8557**	**1318.74**	**313.75**	**10.18**
Ego-Exo4D										
WAN VACE (Jiang et al., 2025)	12.67	0.4161	0.5981	0.8251	0.0362	0.2214	0.8003	6810.62	2163.18	**74.00**
EgoWorld (Park et al., 2026)	11.05	0.3410	0.7186	0.7220	0.0227	0.2349	0.7911	–	–	–
Vista4D (Lin et al., 2026b)	10.46	0.2684	0.7348	0.7459	0.0154	0.2571	0.8067	20074.43	2293.86	120.68
EgoX (Kang et al., 2026)	13.14	0.4474	0.5258	0.8563	0.0703	0.1842	0.8298	5773.68	1928.07	75.55
Ours	**14.11**	**0.4690**	**0.5035**	**0.8665**	**0.0723**	**0.1745**	**0.8341**	**5208.17**	**1749.36**	93.81

extract object masks using SAM 2 (Ravi et al., 2025), and report mask IoU (mIoU), normalized object center error (Center-Err.), and object-crop CLIP similarity (CLIP-O). For *hand consistency*, we extract hand using (Potamias et al., 2025) and report MPJPE, per-frame Procrustes-aligned MPJPE (PA-MPJPE), and sequence-level world-aligned MPJPE (WA-MPJPE) in millimeters. Detailed definitions of the metrics are provided in the supplementary materials.

Implementation Details. We initialize the video DiT from Wan2.1-I2V-14B (Wan et al., 2025), which consists of 40 blocks of width 5120. We fine-tuned the model using LoRA Hu et al. (2022), rank=256, on 8 H200 GPUs, with the AdamW optimizer, learning rate 2×10^{-5}, betas $(0.9, 0.95)$, and weight decay 10^{-4}. We jointly optimize the LoRA parameters, the dual-branch HOI residual adapter, and the cross-attention gates, while keeping the remaining pretrained parameters frozen. Full architecture and configuration details are given in the supplementary materials.

Baselines. We compare our approach with recent exocentric-to-egocentric video generation methods, including EgoWorld (Park et al., 2026) and EgoX (Kang et al., 2026), as well as general view-controllable video generation baselines, WAN VACE (Jiang et al., 2025) and Vista4D (Lin et al., 2026b). Since Vista4D provides inference-only code and EgoWorld does not include the complete training pipeline required for fine-tuning, we evaluate both methods using their publicly available inference implementations. For WAN VACE and EgoX, we fine-tune the released models using the same training configuration as ours for a fair comparison.

4.2 Main Results

We evaluate exocentric-to-egocentric video generation on ARCTIC-HOI and Ego-Exo4D, with quantitative results reported in Tab. 1 and qualitative comparisons shown in Fig. 4.

Quantitative Comparison. As shown in Tab. 1, Exo2EgoHOI consistently improves visual fidelity, object consistency, and hand consistency on ARCTIC-HOI. Compared with the strongest baselines, LPIPS decreases from 0.4324 to 0.3860, while object mIoU increases from 0.2933 to 0.3881. More importantly, MPJPE decreases from 2019.81 to 1318.74 mm and PA-MPJPE from 20.36 to 10.18 mm, corresponding to reductions of 34.7% and 50.0%, respectively. These improvements indicate that our explicit HOI conditioning better preserves both manipulated objects and hand configurations across viewpoints. On Ego-Exo4D, Exo2EgoHOI similarly improves most visual and object metrics, increasing PSNR from 13.14 to 14.11 dB and improving all three object-consistency metrics over EgoX. MPJPE and WA-MPJPE are also reduced to 5208.17 and 1749.36 mm, respectively, although PA-MPJPE remains higher than WAN VACE and EgoX. Overall, these results demonstrate substantial gains in object consistency and HOI preservation while maintaining strong visual fidelity across both datasets. EgoWorld's hand metrics are marked as "–" because its hand estimator component, WiLoR (Potamias et al., 2025), could not reliably recover hand meshes from its generated videos.

Qualitative Comparison. Fig. 4 compares generated egocentric views together with their estimated hand meshes. Existing methods can produce plausible egocentric results but often alter object appearance, placement, or hand configuration under large viewpoint changes. In the hinged-object example, our generated video better preserves the raised lid and circular interior of the target object,

Table 2: **Ablation studies on the ARCTIC-HOI dataset.** "#8 Vanilla CA" uses CLIP to encode the entire anchor frame for cross-attention without decomposition. Best results are shown in **bold** and second-best results are underlined.

Methods	HOI Priors			DGCA			Evaluation Metrics						
	Scene	Hand	Inter.	Global	Local	Gate	LPIPS↓	CLIP-I↑	mIoU↑	CLIP-O↑	MPJPE↓	WA-MPJPE↓	PA-MPJPE↓
#0 Baseline	✗	✗	✗	✗	✗	✗	0.5577	0.8253	0.1148	0.8056	2631.66	371.25	41.42
#1 *w/o* All Priors	✗	✗	✗	✓	✓	✓	0.5351	0.8449	0.1378	0.8255	2254.26	346.98	21.11
#2 *w/* Scene	✓	✗	✗	✓	✓	✓	0.4435	0.9008	0.3104	0.8391	1978.57	327.60	14.30
#3 *w/* Scene & Hand	✓	✓	✗	✓	✓	✓	0.4063	0.9108	0.3323	0.8421	1570.51	315.28	13.33
#4 *w/o* Decomp. CA	✓	✓	✓	✗	✗	✗	0.3953	0.8853	0.3023	0.8198	2327.34	398.94	16.02
#5 *w/* Global	✓	✓	✓	✓	✗	✗	0.4258	0.9057	0.3245	0.8405	1805.71	323.42	14.91
#6 *w/* Local	✓	✓	✓	✗	✓	✗	0.4326	0.9025	0.3559	0.8464	1845.20	321.22	13.30
#7 *w/o* Gate	✓	✓	✓	✓	✓	✗	0.4005	0.9156	0.3354	0.8479	1638.83	322.44	13.61
#8 Vanilla CA	✓	✓	✓	–	–	–	0.4167	0.9108	0.3618	0.8488	1598.81	320.69	11.39
#9 Ours (Full)	✓	✓	✓	✓	✓	✓	**0.3860**	**0.9175**	**0.3881**	**0.8557**	**1318.74**	**313.75**	**10.18**



Figure 4: **Qualitative comparison on ARCTIC-HOI dataset.** Compared with existing methods, Exo2EgoHOI better preserves the target-view object appearance and hand-object configuration under large viewpoint changes. Insets visualize hand meshes recovered from the generated egocentric frames for comparison with the reference.

while in the scissors example ours more closely reproduces the target-view tool placement and bimanual configuration. Similar improvements are observed in the remaining examples, where Exo2EgoHOI better maintains object consistency and HOI configurations during video generation.

4.3 Ablation Studies

We ablate the key designs of Exo2EgoHOI on ARCTIC-HOI in Tab. 2. Fig. 5 provides qualitative comparisons and analyzes the effect of using different numbers of exocentric training views.

Unified 4D HOI Prior. We first study the complementary roles of scene, hand, and interaction priors. Without geometric priors (#1), the model obtains an mIoU of 0.1378 and MPJPE of 2254.26 mm. Introducing the scene prior (#2) substantially increases mIoU to 0.3104 and reduces MPJPE to 1978.57 mm, demonstrating the importance of target-view scene geometry. Adding the articulated hand prior (#3) further reduces MPJPE to 1570.51 mm and PA-MPJPE from 14.30 to 13.33 mm, showing that explicit hand geometry provides finer constraints beyond scene-level conditioning. Finally, adding the interaction prior while keeping DGCA unchanged (#3→#9) increases mIoU from 0.3323 to 0.3881 and reduces MPJPE and PA-MPJPE to 1318.74 and 10.18 mm, respectively. These results show that scene layout, hand articulation, and hand-object relations provide complementary levels of geometric guidance, with explicit relational cues further improving both object and hand consistency.

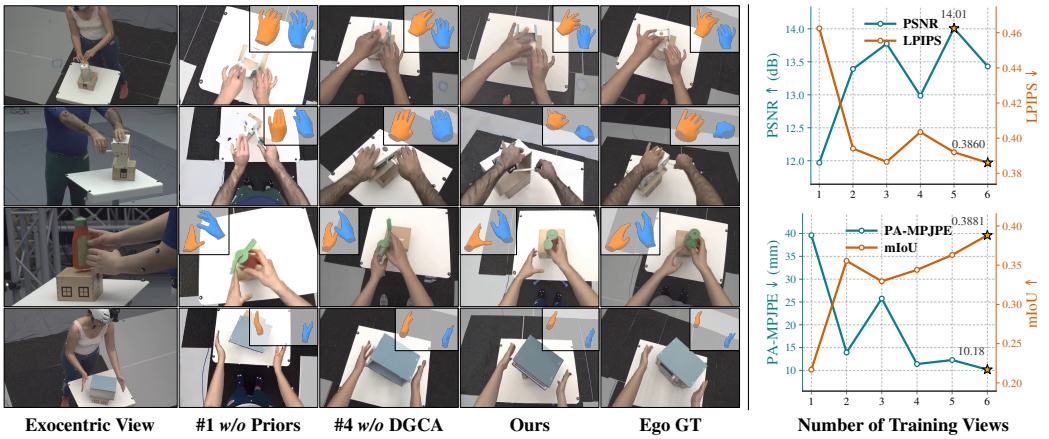


Figure 5: **Ablation studies on the ARCTIC-HOI dataset.** Left: qualitative comparison of different configurations. Right: visual fidelity (PSNR and LPIPS; top) and hand/object consistency (PA-MPJPE and mIoU; bottom) with different numbers of exocentric training views.

Decomposed Gated Cross-Attention. We next analyze the components of DGCA for preserving object consistency while retaining all HOI priors (#4–#9). Compared with no DGCA (#4), introducing either the global (#5) or local (#6) branch improves object and hand consistency. The global branch achieves higher CLIP-I, while the local branch provides better mIoU and PA-MPJPE, supporting their complementary roles in providing semantic and fine-grained appearance anchors for object consistency. Combining both branches with fixed fusion (#7) further improves CLIP-I and MPJPE, while replacing fixed weights with learned per-block gates (#9) increases mIoU from 0.3354 to 0.3881 and reduces PA-MPJPE from 13.61 to 10.18 mm. Moreover, using global and local features without object-background decomposition (#8) degrades mIoU from 0.3881 to 0.3618 and PA-MPJPE from 10.18 to 11.39 mm. Together, these results validate both regional decomposition and adaptive global-local fusion in DGCA. Qualitative examples in Fig. 5 show consistent trends, with the full model better preserving object appearance, state, and hand-object configurations across the generated sequence.

Multi-View Exocentric Training. We further study robustness to source viewpoints by training with one to six exocentric cameras under the same number of optimization steps. As shown in Fig. 5, increasing viewpoint diversity generally improves performance, although the trend is not strictly monotonic. PSNR peaks at 14.01 dB with five views, while six-view training achieves the best LPIPS (0.3860), mIoU (0.3881), and PA-MPJPE (10.18 mm). These results suggest that diverse exocentric training viewpoints improve generalization to viewpoint changes, with performance largely saturating around five to six views.

5 Conclusion

We present Exo2EgoHOI, an HOI-aware framework for exocentric-to-egocentric video generation that aims to faithfully preserve demonstrated hand-object interactions across large viewpoint changes. Our key idea is to complement scene-level geometric conditioning with explicit interaction guidance and object-centric anchoring. The unified 4D HOI prior jointly models scene structure, articulated hands, and hand-object relations to preserve fine-grained interaction, while Decomposed Gated Cross-Attention anchors object-specific semantic and appearance cues throughout egocentric video generation to preserve object consistency across viewpoints. Experiments on ARCTIC-HOI and Ego-Exo4D show that Exo2EgoHOI better preserves manipulated objects and hand configurations under large viewpoint changes. On ARCTIC-HOI, it improves object mIoU by 32.3% and reduces PA-MPJPE by 50.0% relative to the respective best baselines.

Limitations and future work. Our framework relies on estimated depth, hand geometry, and object segmentation, whose errors may propagate into the reconstructed 4D HOI prior and affect generation quality. Future work will explore more robust interaction reconstruction and the use of generated egocentric videos as human demonstrations for downstream embodied learning and robot manipulation.

AI USE STATEMENT

ChatGPT and Codex were used during manuscript preparation to support language editing, including grammar correction, sentence refinement, readability improvement, and consistency of technical terminology. All AI-assisted edits were carefully reviewed and revised by the authors to ensure that they faithfully reflect the intended scientific claims, methodology, and experimental results. The authors assume full responsibility for the final manuscript.

ETHICS STATEMENT

The authors acknowledge and adhere to the ICLR Code of Ethics. The research presented in this paper, as well as our participation in the submission and review process, follows the ethical principles and professional standards established by ICLR.

REPRODUCIBILITY STATEMENT

To ensure reproducibility and verification of our work, we include the implementation details and evaluation procedures in our appendix and will make our source code publicly available upon acceptance.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *International Conference on Learning Representations*, volume 2025, pp. 66712–66737, 2025.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. ISSN 0893-6080.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18973–18990, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d:

Understanding skilled human activity from first-and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025.

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11116–11126, 2026.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233, 2025.

Dayou Li, Lulin Liu, Bangya Liu, Shijie Zhou, Jiu Feng, Ziqi Lu, Minghui Zheng, Chenyu You, and Zhiwen Fan. Egocentric world model for photorealistic hand-object interaction synthesis. In *European Conference on Computer Vision*, 2026.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: recovering the visual space from any views. In *International Conference on Learning Representations*, 2026a.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32671–32682, 2026b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Z Shou. Exocentric-to-egocentric video generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136149–136172, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024b.

Xueyi Liu and Li Yi. GeneOH Diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *International Conference on Learning Representations*, volume 2024, pp. 3652–3686, 2024.

Zuxing Lu, Hongjia Zhai, Guanzhi Wang, Huajian Zeng, Jiaqi Yang, Jingyu Liu, Lei Cheng, Yuantai Zhang, Yuheng Qiu, Zezhou Cheng, et al. World-action models for robot learning and control: A survey. *arXiv:2609.16074*, 2026.

Mohammad Mahdi, Yuqian Fu, Nedko Savov, Jiancheng Pan, Danda Pani Paudel, and Luc Van Gool. Exo2egosyn: Unlocking foundation video generation models for exocentric-to-egocentric video synthesis. *arXiv:2511.20186*, 2025.

Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *IEEE International Conference on Robotics and Automation*, pp. 3400–3407, 2011.

Junho Park, Andrew Sangwoo Ye, and Taein Kwon. Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. In *International Conference on Learning Representations*, 2026.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *International Conference on 3D Vision (3DV)*, pp. 933–943, 2024.

InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv:2604.07209*, 2026.

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:2503.20314*, 2025.

Shengze Wang, Michael Stengel, Tianye Li, Seonwook Park, Amrita Mazumdar, Koki Nagano, Alex Trevithick, and Shalini De Mello. Grounded-exo2ego: Structured semantic grounding for robust exocentric-to-egocentric video generation. *arXiv:2608.20534*, 2026a.

Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2world: Autoregressive egocentric interaction generation via free-space hand gestures. *arXiv:2602.09600*, 2026b.

Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. Humanego: Zero-shot robot learning from minutes of human egocentric videos, 2026c.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. Egovla: Learning vision-language-action models from egocentric human videos, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2025.

Huajian Zeng, Lingyun Chen, Jiaqi Yang, Yuantai Zhang, Fan Shi, Peidong Liu, and Xingxing Zuo. Flowhoi: Flow-based semantics-grounded generation of hand-object interactions for dexterous robot manipulation. *arXiv:2602.13444*, 2026a.

Huajian Zeng, Chaohua Yao, Yuantai Zhang, Jiaqi Yang, Rolandos Alexandros Potamias, and Xingxing Zuo. Stablehand: Quality-aware flow matching for world-space dual-hand motion estimation from egocentric video. *arXiv:2605.18553*, 2026b.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. Egoscale: Scaling dexterous manipulation with diverse egocentric human data, 2026.

Jiaming Zhou, Qihang Zhang, Gangwei Xu, Cunxin Fan, Yujie Zhao, Ruilin Wang, Yiming Luo, Shuai Yang, Xing Zhu, Yujun Shen, Junwei Liang, and Yinghao Xu. Zero-wam: In-context world-action modeling from human videos for open-ended task generalization, 2026.

A Appendix

This supplementary material provides additional implementation, evaluation, and qualitative details for Exo2EgoHOI. We first describe the model architecture, HOI adapter, global-local anchor construction, and training and inference settings. We then detail the construction of ARCTIC-HOI and the preprocessing pipeline for scene geometry, object masks, and hand reconstruction. Next, we provide complete definitions of the visual fidelity, object consistency, and hand consistency metrics used in the main paper. Finally, we present additional qualitative comparisons on ARCTIC-HOI and Ego-Exo4D, followed by representative failure cases and an analysis of the current limitations.

A.1 Architecture and Implementation Details

Backbone and latent layout. We use Wan2.1-I2V-14B-480P (Wan et al., 2025) as our backbone, which contains 40 transformer blocks, hidden width 5120, 40 attention heads, and latent patch size $(1, 2, 2)$. Our 36 input channels comprise the current video latent (dim=16), conditioning mask (dim=4), and visual condition (dim=16). The frozen causal Wan VAE has encoder widths $(96, 192, 384, 384)$ and two residual blocks per stage. Spatial compression is eight; temporal compression maps T frames to $(T-1)/4+1$. Exocentric, ego-target, and rendered scene-prior videos are encoded separately before widthwise latent concatenation. Exocentric 448×784 and ego 448×448 images produce a combined $13 \times 56 \times 154$ latent grid, or $13 \times 28 \times 77$ transformer tokens. Only ego latents receive noise and supervision during the training process.

Operation	Channels	Kernel	Stride	Output (T, H, W)
Branch Conv 1–3	$C \to 16 \to 16 \to 16$	$(3, 3, 3)$	$(1, 1, 1)$	$(49, 448, 448)$
Branch Conv 4	$16 \to 16$	$(3, 3, 3)$	$(1, 2, 2)$	$(49, 224, 224)$
Branch Conv 5	$16 \to 16$	$(3, 3, 3)$	$(2, 2, 2)$	$(25, 112, 112)$
Branch Conv 6	$16 \to 16$	$(3, 3, 3)$	$(2, 2, 2)$	$(13, 56, 56)$
Concatenate + fuse	$16 + 16 \to 16$	$(1, 1, 1)$	$(1, 1, 1)$	$(13, 56, 56)$
Patch projection	$16 \to 5120$	$(1, 2, 2)$	$(1, 2, 2)$	$(13, 28, 28)$

Table 3: HOI adapter layers; $C = 3$ for hand mesh prior and $C = 5$ for interaction prior. Branch convolutions use padding one and SiLU, fusion and projection use zero padding.

Dual-branch HOI adapter. Independent branches encode hand prior and the five-channel interaction prior directly in pixel space (see Tab. 3 for detailed architecture). Hand prior retains $[-1, 1]$ scaling and the interaction prior retains its geometric normalization. Every branch convolution has bias and SiLU (Elfwing et al., 2018), with symmetric padding. The final projection's weights and biases are zero-initialized. The resulting residual is initially zero and is added once to the ego token grid immediately after input patchification.

Global-Local Anchor Construction. Frozen CLIP-ViT-H/14 (Radford et al., 2021) resizes images to 224^2 without center cropping and supplies penultimate-layer features: one $[\texttt{CLS}]$ plus 256 patch

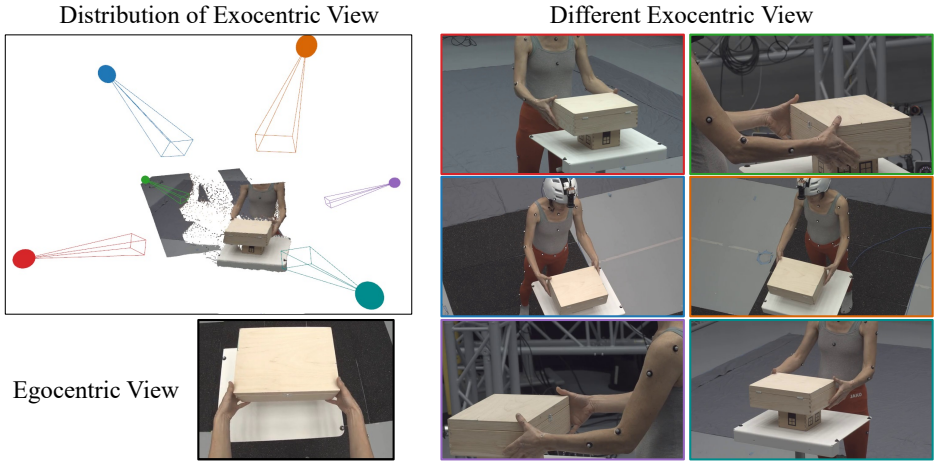


Figure 6: Left: reconstructed 3D scene and camera configuration, with the corresponding egocentric observation shown below. Colored camera frustums denote different exocentric viewpoints. Right: corresponding images of the same scene observed from these exocentric cameras.

tokens, each 1280-dimensional. The object crop expands the tight mask box by 20% on each side and pads to a square with RGB value 0.5. The background image replaces object pixels with 0.5. Their [CLS] tokens form the global reference. A shared pretrained projector maps features to width 5120. Each block shares image key/value projections across global and local contexts.

Training and inference settings. We jointly train the LoRA parameters, adapter, and gates, while freezing all other pretrained weights. LoRA adapts the query, key, value, and output projections in self- and cross-attention, with rank and alpha both set to 256. The flow loss uses timesteps sampled uniformly from a 1000-step Euler training schedule and is computed as the egocentric latent velocity MSE. We use AdamW with learning rate 2×10^{-5}, betas $(0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 10^{-4}. The learning rate remains constant after 100 warm-up steps, and the gradient clipping threshold is 1. Training uses four processes, a per-process batch size of 1, gradient accumulation of 1, and seed 42. Inference uses second-order UniPC (Zhao et al., 2023) with flow shift 3, 50 sampling steps, and guidance scale 5 to generate 49-frame videos at 30 fps.

A.2 ARCTIC-HOI Benchmark Construction

ARCTIC-HOI Dataset Construction. As illustrated in Fig. 6, we construct paired exocentric–egocentric video clips from the ARCTIC (Fan et al., 2023) sequences. We designate camera 0 as the egocentric view and cameras $\{1, 3, 4, 5, 6, 8\}$ as exocentric views. Cameras 2 and 7 are excluded because their viewpoints deviate substantially from the designated egocentric view. After removing the first and last 40 frames of each sequence, we independently sample four non-overlapping 49-frame intervals for each exocentric-egocentric camera pair. This procedure yields approximately 7,000 training clips and 180 test clips, with 30 test clips for each exocentric viewpoint. To further evaluate robustness in unconstrained scenarios, we additionally collect 60 paired exocentric–egocentric clips from in-the-wild human manipulation videos. For these sequences, the relative camera poses between the first- and third-person views are calibrated using AprilTag (Olson, 2011) placed on the tabletop. These videos cover more diverse environments, objects, and manipulation patterns beyond the controlled ARCTIC setting. Combining the ARCTIC-derived split with these in-the-wild sequences, our ARCTIC-HOI benchmark contains approximately 7,000 training clips and 240 test clips in total. We use Depth Anything 3 (DA3) (Lin et al., 2026a) (`DA3NESTED-GIANT-LARGE-1.1`) to reconstruct scene geometry offline. Importantly, no egocentric RGB images are used for depth estimation. Our video generation takes only a single exocentric clip together with the target camera trajectory. For each selected source view, we back-project its RGB-D observation into a point cloud and render only this source-view point cloud into the target egocentric camera. Rendering uses a depth buffer, point size 7, and $2\times$ supersampling, with uncovered regions set to black.

Object Masks and Hand Reconstruction. We obtain object and hand masks using GroundingDINO Swin-T OGC (Liu et al., 2024b) followed by SAM 2.1 Hiera-Large (Ravi et al., 2025). Objects and hands are detected separately using the prompts `object.` and `hand.`, with box and text thresholds of 0.35 and 0.25, respectively. Masks detected under the same prompt are merged, and segmentation is performed independently for each frame without temporal mask propagation. For hand reconstruction, WiLoR (Potamias et al., 2025) predicts 778-vertex MANO (Romero et al., 2017) meshes with a confidence threshold of 0.3, an IoU threshold of 0.5, and crop rescaling factor 2.0. We align each reconstructed hand mesh to the observed scene depth using valid depth measurements within the corresponding hand mask. Depth alignment preserves the estimated hand articulation, orientation, and scale, while hands without sufficient depth support are discarded. Finally, egocentric hand priors are rendered using fixed colors for the left and right hands with hand-only visibility, without applying object-depth occlusion.

A.3 Evaluation metrics

A.3.1 Visual fidelity

We evaluate visual fidelity using PSNR, SSIM, LPIPS, and CLIP-I, with scores computed per frame before averaging. PSNR and SSIM use RGB images with a data range of 255 and uniform 7×7 windows for SSIM. LPIPS uses pretrained AlexNet, with inputs normalized to $[-1, 1]$ and no additional resizing. CLIP-I measures cosine similarity between ℓ_2-normalized image features from `openai/clip-vit-base-patch32`. Its preprocessing consists of bicubic short-side resizing to 224, and a 224×224 center crop.

A.3.2 Object consistency

We independently extract object masks from each reference and generated frame using Grounding DINO Swin-T OGC (Liu et al., 2024b) and the SAM 2.1 Hiera-L image predictor (Ravi et al., 2025). We use the prompt `object.`, box and text thresholds of 0.35 and 0.25, and single-mask SAM output. Detector inputs are resized to a short side of 800 with a maximum side of 1333 and use ImageNet normalization.

mIoU and Center-Err. For reference and predicted masks $M_t, \widehat{M}_t$, let $\mathcal{G}$ contain frames with a reference mask and $\mathcal{B}$ contain frames where both masks are present. We compute the following per-video scores:

$$\text{mIoU} = \frac{1}{|\mathcal{G}|} \sum_{t \in \mathcal{G}} \frac{|\widehat{M}_t \cap M_t|}{|\widehat{M}_t \cup M_t|}, \qquad \text{Center-Err.} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \frac{\|c(\widehat{M}_t) - c(M_t)\|_2}{\sqrt{H^2 + W^2}}, \tag{15}$$

where $c(M)$ is the foreground-pixel centroid. mIoU measures mask overlap over reference-present frames, assigning zero to missing predictions and excluding reference-absent frames. Center-Err. measures centroid displacement normalized by the full image diagonal, including padding, and is evaluated only when both masks are present.

CLIP-O. CLIP-O measures object appearance consistency by averaging CLIP image-feature cosine similarity over $\mathcal{B}$, using the same encoder and preprocessing as CLIP-I. Object crops are constructed independently from the reference and predicted masks. Each mask's tight bounding box is expanded by $\lceil 0.2w \rceil$ on each horizontal side and $\lceil 0.2h \rceil$ on each vertical side, where w and h are the box width and height. The expanded crop is clipped to the image and centered on a square canvas. Non-mask pixels and square padding are filled with RGB $(127, 127, 127)$. Both CLIP-O and Center-Err. are conditional on both masks being present and do not penalize missing predictions.

A.3.3 Hand consistency

We use pretrained WiLoR (Potamias et al., 2025) to independently estimate 21 hand joints from reference and generated RGB frames. We use the WiLoR detector with confidence threshold 0.3, NMS IoU threshold 0.5, box expansion factor 2.0, and batch size 32. Pre-processing uses normalized 256^2 hand crops and left-hand flipping followed by restoration. For each frame and reference side, we select the highest-confidence reference detection and the highest-confidence same-side prediction. If no same-side prediction exists, we select the hand with the nearest camera-space wrist among all predicted hands. Reference-missing observations are excluded.

Error definition. For associated joints $\widehat{\mathbf{p}}_{t,s,j}, \mathbf{p}_{t,s,j}$ in millimeter, we define

$$e_{t,s}(A) = \frac{1}{21} \sum_{j=1}^{21} \|A(\widehat{\mathbf{p}}_{t,s,j}) - \mathbf{p}_{t,s,j}\|_2, \qquad A(\mathbf{p}) = aR\mathbf{p} + \mathbf{b}. \tag{16}$$

The three hand metrics differ in their alignment A:

- *MPJPE* uses the identity transformation A and measures joint errors directly in camera coordinates, without wrist centering.
- *PA-MPJPE* fits a separate similarity transformation A for each frame and reference side, allowing scale, rotation, and translation alignment independently at each frame.
- *WA-MPJPE* fits one similarity transformation A per video and reference side using all associated frames, including wrist-fallback matches. The same transformation is applied throughout the sequence. This aligns WiLoR camera-coordinate sequences and does not recover a physical world coordinate system.

For PA-MPJPE and WA-MPJPE, alignment minimizes squared joint residuals via SVD over scale, proper rotation without reflection, and translation. The reported errors in Eq. 16 are mean Euclidean distances in millimeters.

A.4 Additional Qualitative Results

We provide additional qualitative results to further evaluate Exo2EgoHOI across different datasets, viewpoints, and manipulation scenarios. In addition to visual comparisons of the generated egocentric videos, we analyze how well different methods preserve the demonstrated hand configurations and manipulated objects under large exocentric-to-egocentric viewpoint changes.

On ARCTIC-HOI (see Fig. 7), we provide additional comparisons with existing methods and overlay the reconstructed 3D hand meshes on the generated egocentric frames. We further color-code the per-vertex hand errors with respect to the egocentric ground truth, providing a more direct visualization of hand-pose consistency beyond image-level appearance. These examples cover diverse hand configurations, object geometries, and bimanual interactions, and complement the quantitative hand-consistency metrics reported in the main paper.

We also present additional results on Ego-Exo4D (see Fig. 8), which contains more diverse real-world environments, camera configurations, and manipulation activities. These examples provide a complementary evaluation of cross-view generation beyond the relatively controlled ARCTIC-HOI setting, particularly in terms of scene reconstruction, manipulated-object consistency, and hand-object configurations.

Finally, we show representative failure cases (see Fig. 9) of Exo2EgoHOI. Severe occlusion of the hands or manipulated objects can provide insufficient geometric evidence for reconstructing the underlying interaction, resulting in inaccurate target-view generation. Complex articulated or thin-structured objects, such as scissors, are also challenging because their fine-scale geometry and articulation states are difficult to recover reliably from exocentric observations. These cases highlight the dependence of our framework on accurate upstream geometry and interaction reconstruction, and motivate more robust HOI reconstruction for future work.

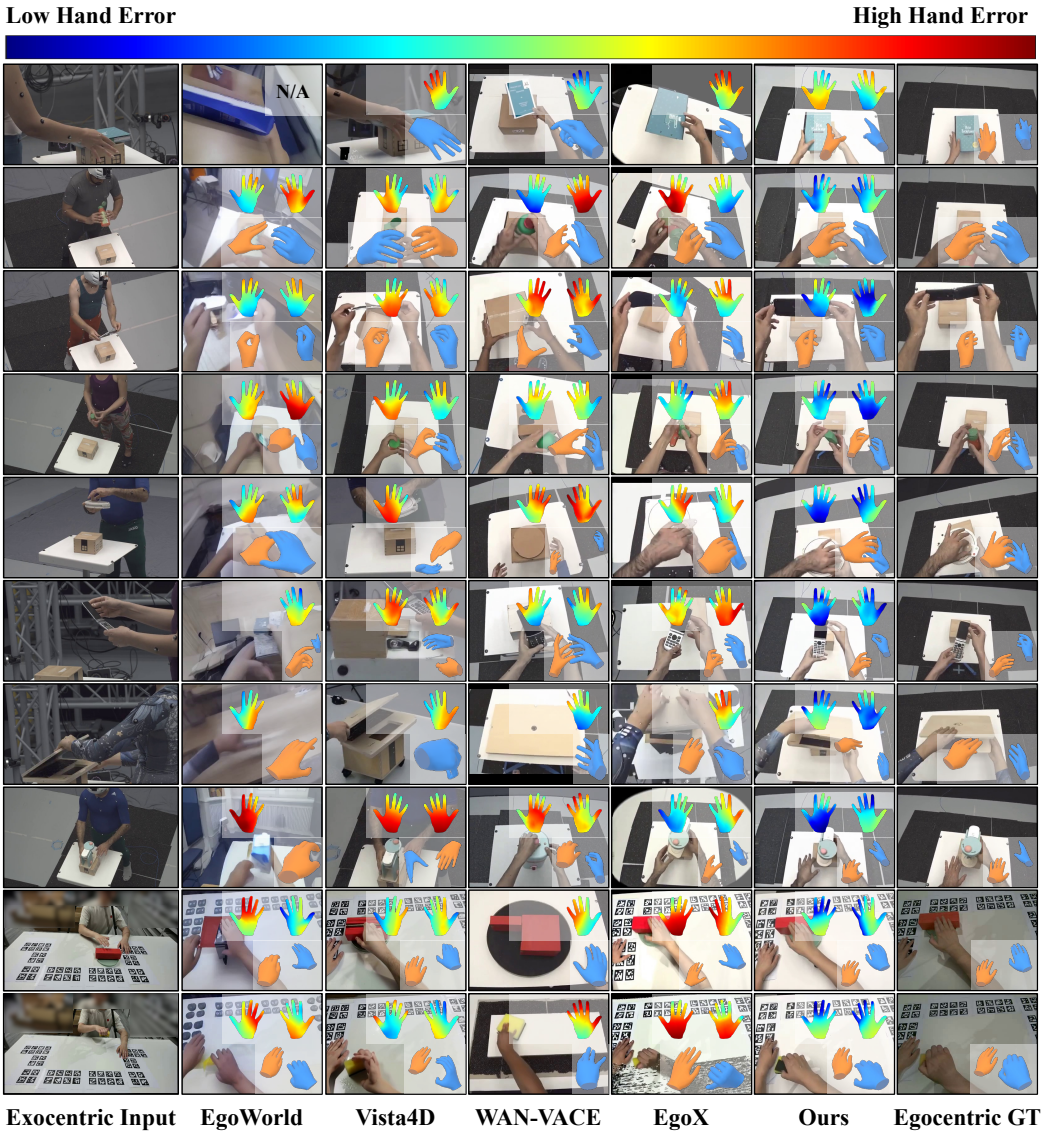


Figure 7: **Additional qualitative comparison on ARCTIC-HOI.** We compare egocentric videos generated by different methods and overlay their reconstructed 3D hand meshes. Hand-mesh vertices are color-coded by errors relative to the egocentric ground truth, ranging from blue (low error) to red (high error); *N/A* indicates failed hand reconstruction. The visualization highlights differences in target-view hand configuration and hand-object alignment across methods.

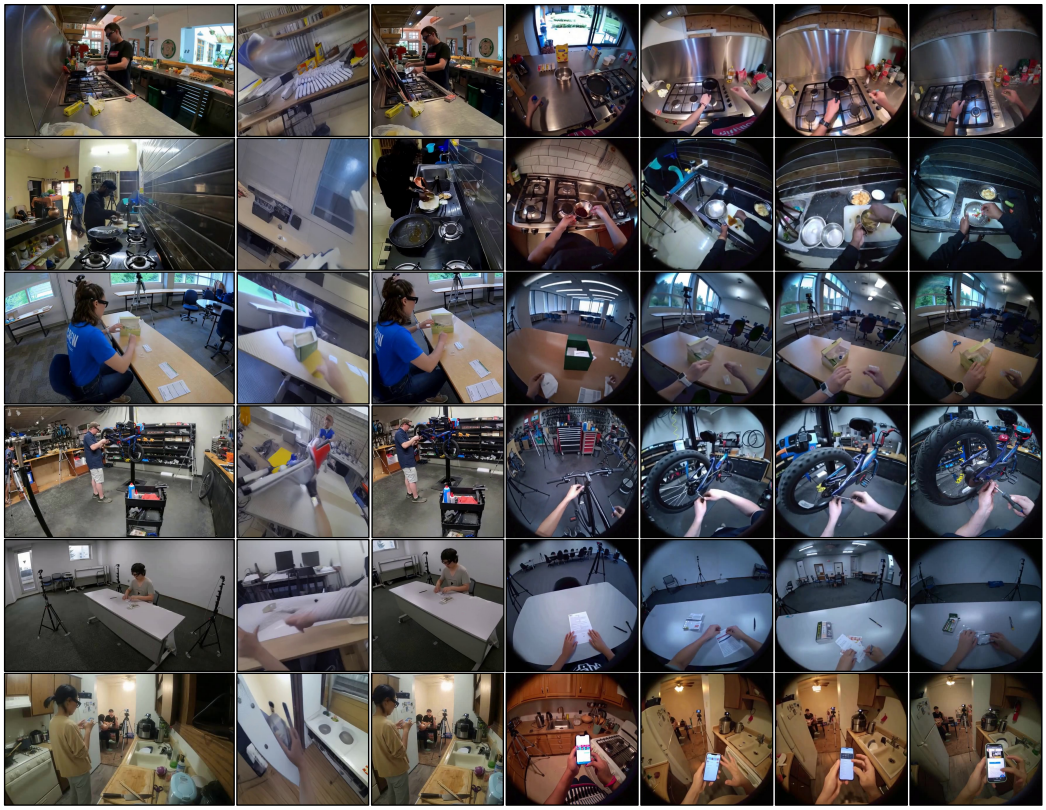


Figure 8: **Additional qualitative comparison on Ego-Exo4D.** We compare egocentric videos generated from exocentric inputs across diverse real-world scenes and manipulation activities. The results illustrate differences among methods in cross-view scene reconstruction, manipulated-object consistency, and hand-object configurations under substantial viewpoint changes.

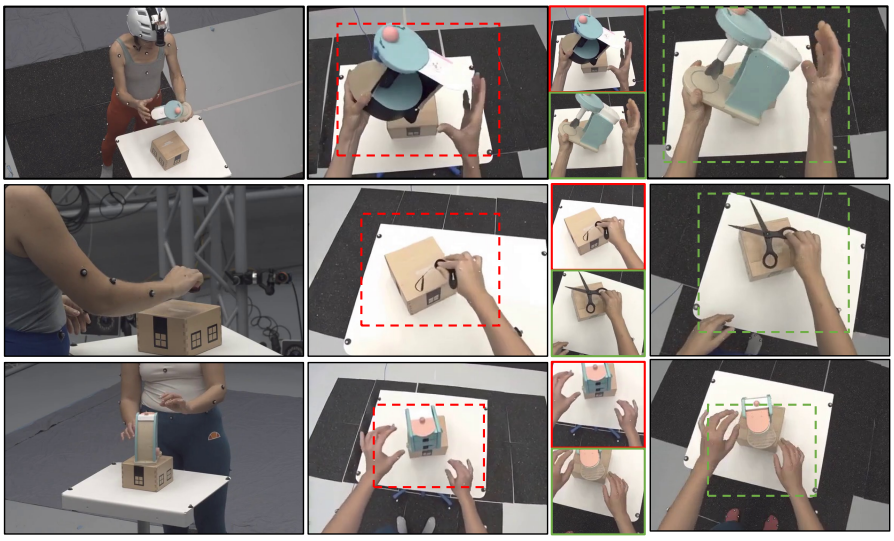


Figure 9: **Representative failure cases of Exo2EgoHOI.** Severe occlusion can lead to incomplete hand-object reconstruction, while complex articulated or thin-structured objects remain difficult to model accurately. Insets highlight discrepancies in object geometry, articulation state, and hand-object configuration compared with the egocentric ground truth.