4D-HOF: HAND-OBJECT FLOW MATCHING FOR FEED-FORWARD 4D INTERACTION RECONSTRUCTION

Anonymous authors
Paper under double-blind review



Figure 1: **4D Hand-Object Flow Matching (4D-HOF)**. Given an in-the-wild monocular video of a hand interacting with an unseen object, 4D-HOF learns a flow matching model for accurate and efficient 4D hand-object interaction reconstruction. 4D-HOF generalizes across diverse and challenging scenarios, including ambiguous depth (headset), occlusion (kettle), complex articulated hand poses (bar), and large object rotation (camera). Moreover, the running time of 4D-HOF is substantially faster than prior optimization-based methods due to our feed-forward design.

ABSTRACT

Existing methods for 4D hand-object reconstruction often rely on costly per-sequence optimization, while generative approaches typically synthesize interactions from random noise, which can lead to unstable interaction prediction. We introduce 4D-HOF, a feed-forward framework that reconstructs 4D hand-object interactions from coarse but informative estimates produced by vision foundation models. Concretely, we learn a conditional flow matching model that transports foundation-model-derived hand-object states toward an interaction manifold, allowing the model to correct errors in translation, rotation, and alignment in a feed-forward manner. A key advantage of our generative formulation is that it naturally enables test-time guidance within the transport process. Rather than applying a separate post-hoc optimization after reconstruction, we directly steer the evolving generative states using physical interaction constraints and observed 2D evidence, allowing the reconstruction to be refined as part of the generative process itself. By training the generative model on diverse datasets, 4D-HOF generalizes robustly to challenging in-the-wild scenarios. Experiments on out-of-domain benchmarks show that 4D-HOF achieves state-of-the-art performance, producing more stable and accurate 4D hand-object reconstructions.

1 INTRODUCTION

Reconstructing 4D hand-object interaction (HOI) from monocular video is a fundamental problem for understanding human behavior and enabling a variety of applications such as robot learning (He

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.

et al., 2025; Zhong et al., 2025; Paliwal et al., 2026), extended reality (Marchand et al., 2015), and human-computer interaction (Pavlovic et al., 2002). Despite recent progress (Fan et al., 2024; Wang et al., 2025b; Shi et al., 2026; Xu et al., 2026), recovering accurate hand-object geometry and interaction remains challenging due to diverse object geometry and pose, severe occlusion, and the inherent ambiguity of monocular observations.

To address these challenges, existing approaches mainly fall into two lines. Optimization-based methods (Fan et al., 2024; Wang et al., 2025b; Shi et al., 2026; Aytekin et al., 2026b; Zhu et al., 2025) formulate HOI reconstruction as a per-sequence fitting problem, enforcing interaction constraints through manually designed objectives such as contact, rendering, and interaction stability. While effective, these methods are computationally expensive and often brittle under imperfect initialization or occlusion. Generative methods (Ye et al., 2023; 2024; Christen et al., 2024; Li et al., 2025; Ye et al., 2026) learn priors over HOI distributions from large-scale data, yet often start from random noise or rely on category-level priors with closed-set vocabularies, making it difficult to preserve instance-specific geometry and pose cues. This raises a key question: can we retain and learned priors of generative modeling while starting from informative observations rather than random noise?

Vision foundation models provide a natural starting point. Recent large-scale pretrained models can estimate hand pose, object geometry, depth, and scene structure from monocular inputs (Carion et al., 2026; Oquab et al., 2023; Piccinelli et al., 2025; Lin et al., 2025b; Wen et al., 2024; Pavlakos et al., 2024; Potamias et al., 2025). These models produce coarse yet informative estimates that capture rich geometric cues. Indeed, recent works have begun to incorporate foundation models into HOI reconstruction (Liu et al., 2025b; Wang et al., 2026; Lin et al., 2025a; Xu et al., 2026). However, they typically treat these estimates as initialization and rely on downstream heuristic optimization to enforce consistency, which reintroduces inefficiency and instability.

In this work, we instead couple foundation-model perception directly with generative modeling. We use them to define the source distribution of a conditional generative transport process. Starting from coarse but informative hand-object states, we learn how these states should evolve toward physically plausible interactions. Specifically, we adopt flow matching and stochastic interpolates (Liu et al., 2022b; Shi et al., 2023; Albergo et al., 2025) to construct a generative bridge that transports the initial estimates toward an HOI manifold. To ground refinement in observed evidence and exploit interaction dynamics, the bridge is conditioned on both the input video frames and the temporal context of the initial 4D hand-object sequence. This formulation enables unified correction of geometry, pose, and alignment in a feed-forward manner, while preserving instance-specific cues from the input. Importantly, the generative formulation also gives us a natural mechanism for test-time control to further enhance fine-grained physical plausibility and consistency with the 2D observation. Our method, termed **4D-HOF**, is a feed-forward framework that takes foundation-model estimates as input and jointly refines hand-object interactions. We summarize our contributions as follows:

- We formulate 4D HOI reconstruction as conditional generative transport from foundation-model-derived hand-object state initialization to physically plausible interactions, avoiding random noise as initialization to flow matching and costly per-sequence optimization.
- We design a conditional flow matching model that jointly refines hand-object geometry, pose, and alignment using visual, geometric, and temporal cues.
- We introduce test-time guidance enabled by the generative transport process, which steers evolving states with image, interaction, and temporal constraints without separate post-hoc optimization.
- Experiments on 4D HOI benchmarks and challenging in-the-wild scenarios demonstrate state-of-the-art performance in hand estimation, object reconstruction, interaction consistency, and temporal stability, while running over 5 times faster than prior methods.

2 Related Work

Hand-Object Reconstruction. Early image-based methods jointly regress hand and object meshes using parametric hand priors and contact constraints (Hasson et al., 2019; Ye et al., 2022). Video-based methods typically optimize each sequence with hand-crafted geometric and interaction objectives (Hasson et al., 2021; Wen et al., 2023; Fan et al., 2024; Wang et al., 2025b; Lin et al., 2025a; Shi et al., 2026; Aytekin et al., 2026a; On et al., 2025; Liu et al., 2025a; Aboukhadra et al., 2026), incurring high computational costs and sensitivity to initialization. Feed-forward approaches accelerate inference (Chen et al., 2026b) but emphasize geometry without explicitly modeling interaction

dynamics. Our method instead learns an interaction prior that maps coarse estimates to physically plausible interactions for feed-forward reconstruction.

Foundation Models for 3D Understanding. Foundation models provide strong priors for segmentation (Carion et al., 2026), visual representations (Oquab et al., 2023), monocular and multi-view geometry (Kong et al., 2026; Piccinelli et al., 2025; Lin et al., 2025b; Wen et al., 2025; 2026; Wang et al., 2025a), object pose (Wen et al., 2024), hand reconstruction (Pavlakos et al., 2024; Potamias et al., 2025), and 3D generation (Chen et al., 2026a; Xiang et al., 2026). Recent HOI reconstruction pipelines use these models for initialization (Wang et al., 2026; Lin et al., 2025a; Xie et al., 2026), but independently estimated components often have inconsistent scale, depth, and orientation. Existing methods resolve these discrepancies through additional optimization (Shi et al., 2026; Aytekin et al., 2026a); we learn to jointly refine the estimates without per-sequence optimization.

Generative Priors for Hand-Object Interaction. Generative models capture HOI distributions through diffusion (Ho et al., 2020; Christen et al., 2024; Li et al., 2025) or language modeling (Huang et al., 2025a), primarily for interaction synthesis. Complementary denoising methods learn spatio-temporal priors to refine noisy HOI trajectories (Zhou et al., 2022; Liu & Yi, 2024; Luo et al., 2024), demonstrating the potential of learned priors for HOI correction. For reconstruction, Diff-HOI (Ye et al., 2023) and G-HOP (Ye et al., 2024) incorporate generative priors, but they rely on noise-driven generation or category-level priors with closed-set vocabulary. Most closely related, CHOIR (Xu et al., 2026) learns a flow-matching prior to correct relative hand-object 1D depth displacement along the viewing ray from initializations, while contact-aware optimization refines the full interaction. Building on flow matching and stochastic interpolants (Liu et al., 2022b; Shi et al., 2023; Albergo et al., 2025), we instead learn a conditional bridge that jointly refines entire HOI states, making the learned prior the primary reconstruction mechanism, complemented by test-time guidance.

Agentic Real-to-Sim. Recent advances in agentic AI motivate its integration into 3D reconstruction pipelines. AgentSTAR (Mazur et al., 2026) couples visual reasoning with numerical optimization in a render-and-compare loop to recover object geometry, articulation, and motion, but does not explicitly model hand–object interactions. Inspired by GPT-6 Astra's capabilities, we conduct pilot experiments on agentic HOI reconstruction, detailed in the Appendix B.4. Our preliminary findings suggest that this approach is promising but incurs substantial runtime and costs. Our proposed feed-forward method offers an efficient reconstruction module that agents could use to accelerate this process.

3 Approach

3.1 Problem Formulation

4D HOI reconstruction. Given an N-frame monocular RGB video $\mathcal{I} = \{\mathbf{I}^i\}_{i=1}^N$, where a hand interacts with an object of arbitrary shape, our goal is to reconstruct the underlying 4D hand-object interaction. We aim to reconstruct the object mesh $\mathbf{m}_o$ with corresponding per-frame 6-DoF poses $\mathbf{o}^i = (\mathbf{R}^o, \mathbf{t}^o)$, where $\mathbf{R}^o \in SO(3)$ and translation $\mathbf{t}^o \in \mathbb{R}^3$. For the hand, we estimate the per-frame MANO (Romero et al., 2022) parameters, consisting of shape $\boldsymbol{\beta}$ and pose $\mathbf{h}^i = (\mathbf{R}^h, \mathbf{t}^h)$. We denote the full interaction trajectory as $\mathcal{S} = \{\mathbf{s}^i\}_{i=1}^N$, where $\mathbf{s}^i = [\mathbf{h}^i, \mathbf{o}^i]$.

Flow matching for HOI reconstruction. Flow matching (Lipman et al., 2022; Liu et al., 2022a) formulates generative modeling as learning a continuous transport from a source distribution $p(\mathbf{x}_0)$ to a target data distribution $p(\mathbf{x}_1)$. Let $\mathbf{x}_0 \sim p(\mathbf{x}_0)$ and $\mathbf{x}_1 \sim p(\mathbf{x}_1)$. A stochastic interpolant defines intermediate states $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1 + \gamma(t)\boldsymbol{\epsilon}$, where $t \in [0,1]$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and $\gamma(t)$ is a noise schedule. Flow matching learns a velocity field $\mathbf{v}_\theta(\mathbf{x}, t)$ such that the trajectory follows $d\mathbf{x}(t)/dt = \mathbf{v}_\theta(\mathbf{x}(t), t)$, transporting samples from $p(\mathbf{x}_0)$ to $p(\mathbf{x}_1)$. In practice, $\mathbf{v}_\theta$ can be a neural network (e.g., DiT (Peebles & Xie, 2023)) trained to match the conditional drift induced by the interpolate via the following objective:

$$\mathcal{L}_\theta = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\boldsymbol{\epsilon}} \left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_t)/(1-t)\|_2^2 \right]. \tag{1}$$

In the context of HOI, the target distribution $p(\mathbf{x}_1)$ corresponds to physically plausible hand-object interactions, which can be learned from large-scale HOI datasets (Wang et al., 2024; Yang et al., 2022; Swamy et al., 2023; Hampali et al., 2020). While several recent works adopt generative models for HOI, they often start from random noise (Ye et al., 2022; 2026), depend on object templates (Ye et al., 2024), or target the task of grasp synthesis (Wu et al., 2026) instead of full 4D HOI reconstruction.

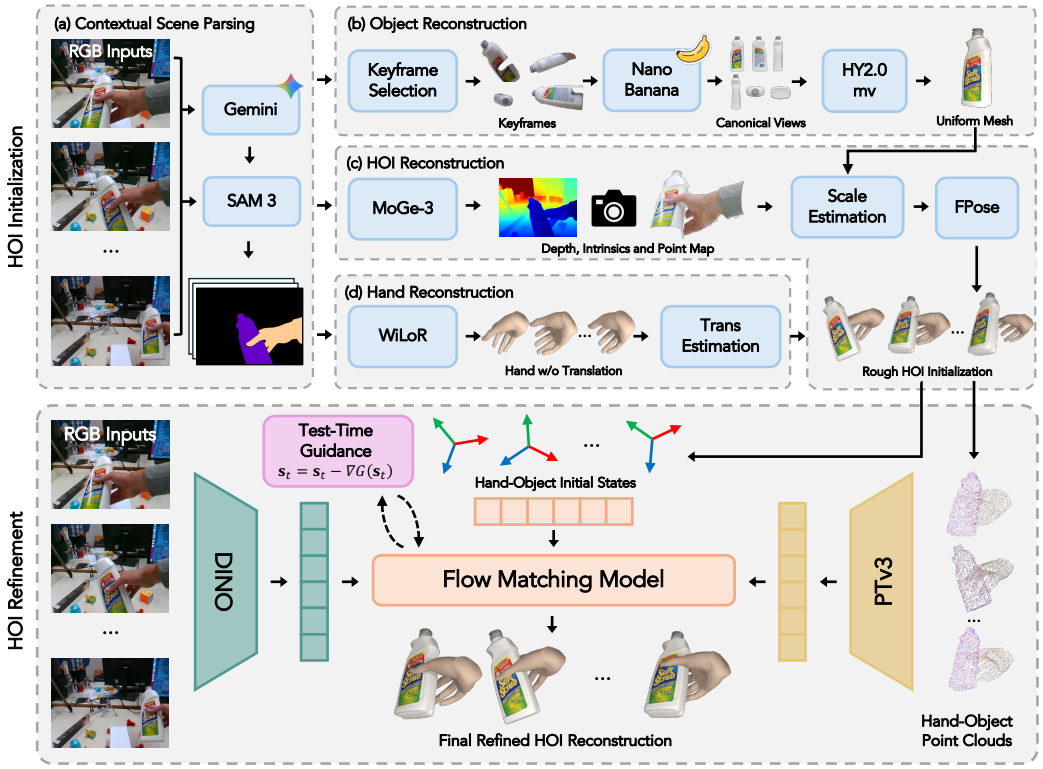

Figure 2: **4D-HOF overview.** Given a monocular video, we construct initial hand-object states by leveraging foundation models, including (a) contextual scene parsing, (b) object reconstruction, (d) hand reconstruction, and (c) depth alignment for recovering metric geometry, producing coarse HOI initialization. We then formulate HOI refinement as a conditional generative bridge matching problem that takes the HOI initialization together with RGB and 3D cues as input and outputs the refined HOI reconstruction. At inference time, test-time guidance adjusts the evolving states to better satisfy physical and image-space constraints.

Bridge matching. A key limitation of vanilla flow matching is that it assumes a fixed Gaussian prior, whose stochastic nature can hinder physical accuracy in many deterministic tasks. To address this, the Bridge Matching (Albergo et al., 2025) framework alleviates this by learning transport between arbitrary distributions π_0 and π_1. Given $\mathbf{x}_0 \sim \pi_0$ and $\mathbf{x}_1 \sim \pi_1$, the trajectory can be written as

$$d\mathbf{x}_t = \mathbf{v}_\theta(\mathbf{x}_t, t)\, dt + \sigma(t)\, d\mathbf{W}_t, \tag{2}$$

where $\mathbf{v}_\theta$ denotes the learnable drift and $\sigma(t)$ controls the noise level. A common instantiation is the Brownian bridge with $\sigma(t) = \sigma\sqrt{t(1-t)}$. The velocity field is trained by matching the drift:

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_t)/(1-t)\|_2^2\right]. \tag{3}$$

When $\sigma(t) = 0$, Eq (3) reduces to deterministic flow matching. By replacing the Gaussian prior with a more plausible initial distribution, it reduces the stochasticity inherent in noise-driven generation. A key remaining question therefore is the choice of the source distribution. In this work, we define it using foundation-model-derived hand and object states, which will be described in the next section.

3.2 Hand and Object States Initialization

The goal is to leverage foundation models to obtain initial hand and object geometry and their per-frame interaction initialization $\mathbf{s}_0^i = [\mathbf{h}_0^i, \mathbf{o}_0^i]$, where i indexes video frames. For brevity, we will omit the frame index i. These estimates provide coarse yet informative hand and object pose, which serve as the starting point of our conditional generative bridge. While foundation models has made significant progress, they treat hand, object, and the context separately and do not consider their intricate interactions. To address this, we introduce the following methods for reconstructing hand, object, and their interaction. Please refer to Appendix for the detailed process.

Contextual scene parsing. We first perform holistic 3D scene understanding to reduce ambiguity in monocular reconstruction and generate outputs for object and hand states estimation. Specifically, for each frame, we obtain the object mask $\mathbf{M}_i$ via vision language grounding (Comanici et al., 2025) and segmentation models (Carion et al., 2026). We also estimate depth $\mathbf{D}_i$ and camera parameters using MoGe-3 (Kong et al., 2026), which provide geometric cues for both hand and object reconstruction. These outputs establish spatial consistency across frames. Figure 2-(a) illustrates the main idea.

Hand reconstruction and state estimation. For hand initialization, we employ WiLoR (Potamias et al., 2025) to estimate per-frame MANO (Romero et al., 2022) parameters and 2D keypoints, with hand bounding boxes extracted from masks. While WiLoR provides reliable hand articulation and rotation estimates, its translation prediction is limited by the weak-perspective camera assumption, which cannot accurately recover metric camera-space depth. To address this, we solve a lightweight closed-form least-squares to estimate the hand translation. Figure 2-(d) shows the pipeline.

Object reconstruction and state estimation. We estimate a canonical object mesh $\mathbf{m}_o$ and temporally consistent per-frame poses $\{\mathbf{o}^i\}_{i=1}^N$. We sample sparse keyframes with diverse viewpoints and reconstruct the mesh in normalized space using pre-trained multi-view generative models (Comanici et al., 2025; Zhao et al., 2025). Compared with single-view generation (Chen et al., 2026a), this reduces ambiguity under occlusion and improves texture quality. We then align the object with the scene using depth cues and an enhanced FoundationPose (Wen et al., 2024), as shown in Figure 2-(b). The resulting states provide plausible structural alignment but remain susceptible to occlusion-induced instability and generative artifacts, motivating subsequent bridge matching refinement.

3.3 4D HOI Flow Matching

After we obtained initialized hand-object states $\mathbf{s}_0$ from the input video sequence, we learn a conditional generative model that transports these states toward physically plausible interactions. We parameterize the conditional velocity field $\mathbf{v}_\theta(\mathbf{s}_t, t)$ using a neural network that takes as input the current state $\mathbf{s}_t$, flow time t, and multi-modal conditions derived from the input video.

Conditional bridge matching. As discussed in Section 3.1, directly applying Eq. (1) to HOI reconstruction is challenging due to the large solution space and ambiguity of monocular observations. To address this issue, we inject an initialization state $\mathbf{s}_0$ through the bridge matching formulation in Eq. (3), which guides the transport process toward physically plausible hand-object configurations.

We further encode geometric structure by representing both the hand and object as point clouds. Specifically, we sample points from the object mesh $\mathbf{m}_o$ and use MANO vertices parameterized by shape coefficients $\boldsymbol{\beta}$ for the hand representation. These point clouds are processed using a Point Transformer V3 backbone (Wu et al., 2024) to capture spatial relationships between the hand and object. To incorporate visual context, we extract image features from each frame using a pretrained DINOv2 encoder (Oquab et al., 2023). For efficient 4D modeling, we aggregate multi-frame features and feed them into a DiT backbone for temporal reasoning and conditional flow prediction. Detailed architecture is shown in Figure S1. The bridge matching objective in Eq. (3) can be written as:

$$\mathcal{L}_{\text{hof}} = \mathbb{E}_{\mathbf{s}_0,\mathbf{s}_1,t}\left[\left\|\mathbf{v}_\theta(\mathbf{s}_t, t, \mathbf{m}_o, \boldsymbol{\beta}, \mathcal{I}) - (\mathbf{s}_1 - \mathbf{s}_t)/(1-t)\right\|_2^2\right], \tag{4}$$

where $\mathbf{m}_o$ denotes the object mesh, $\boldsymbol{\beta}$ the hand shape, and $\mathcal{I}$ the sequence of input images. This objective learns to transport the initialized states $\mathbf{s}_0$ toward the target interaction distribution.

Training objectives. We introduce three auxiliary losses to improve reconstruction quality: 3D position loss $\mathcal{L}_{\text{3D}}$, interaction loss $\mathcal{L}_{\text{inter}}$, and regularization loss $\mathcal{L}_{\text{reg}}$.

We apply regression-based (Pavlakos et al., 2024) 3D position losses $\mathcal{L}_{\text{3D}}$ to the terminal estimate $\mathbf{s}_1$, recovered from the generated trajectory since flow matching predicts velocities rather than states. The 3D loss consists of three components: (1) a hand loss supervising the 3D MANO joints $\mathbf{J}_{\text{3D}}$, (2) an object loss supervising the transformed object point cloud $\mathbf{P}$, and (3) a relative loss supervising the spatial relationship between the hand and the object through their relative Euclidean distances:

$$\mathcal{L}_{\text{3D}} = \|\mathbf{J}_{\text{3D}} - \mathbf{J}^*_{\text{3D}}\|_1 + \|\mathbf{P} - \mathbf{P}^*\|_1 + \|\mathbf{D}(\mathbf{P}, \mathbf{J}_{\text{3D}}) - \mathbf{D}(\mathbf{P}^*, \mathbf{J}^*_{\text{3D}})\|_1, \tag{5}$$

where $\mathbf{P}$ is the point cloud sampled from the transformed object mesh $\mathbf{m}_o$, $\mathbf{D}(\mathbf{P}, \mathbf{J}_{\text{3D}})$ is the distance matrix for each transformed object point-hand joint pair. The asterisk denotes the ground truth.

To encourage physically plausible interactions, we further incorporate contact and penetration-based constraints (Jiang et al., 2021), denoted by $\mathcal{L}_{\text{inter}}$, between the hand mesh and object point cloud:

$$\mathcal{L}_{\text{inter}} = \sum\nolimits_{j\in\Omega} d(p_j, \mathcal{H}) + \sum\nolimits_{j\in\mathcal{P}} d(p_j, \mathcal{H}), \tag{6}$$

where $d(p_j, \mathcal{H})$ is the distance between object point p_j to MANO mesh surface $\mathcal{H}$, Ω is the contact set derived from ground-truth pose and $\mathcal{P}$ is the object point set that penetrates into the hand mesh. The first term encourages contact at ground-truth contact regions, while the second penalizes penetration.

In addition, we empirically observe that the hand estimates provided by the initialization are accurate in most cases. Therefore, the refinement process should make only subtle adjustments to improve physical plausibility, rather than performing unconstrained optimization that may introduce undesirable artifacts. To this end, we introduce a regularization term, $\mathcal{L}_{\text{reg}}$, on hand articulation related channels $\mathbf{v}_{\theta,\text{hand}}$ to preserve the initialized hand motion during refinement:

$$\mathcal{L}_{\text{reg}} = \|\mathbf{v}_{\theta,\text{hand}}(\mathbf{s}_t, t, \mathbf{m}_o, \boldsymbol{\beta}, \mathcal{I})\|_2^2 . \tag{7}$$

The overall training objective combines the bridge matching loss with the auxiliary terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hof}} + \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{reg}}. \tag{8}$$

Test-time guidance. An additional benefit of our bridge matching formulation is that the continuous transport process provides a flexible interface for test-time control. During inference, task-specific constraints can be directly incorporated by steering intermediate states along the learned transport trajectory, without performing expensive per-sequence optimization. This allows us to efficiently impose additional geometric and interaction constraints, such as contact, temporal consistency, and image-space alignment, while preserving the interaction prior learned by the model. Formally, we incorporate the following lightweight test-time guidance $G(\mathbf{s}_t)$ into the bridge matching process:

$$\mathbf{s}_t = \mathbf{s}_t - \nabla_{\mathbf{s}_t} G(\mathbf{s}_t). \tag{9}$$

The guidance function G combines three complementary objectives: silhouette guidance G_{sil} for image alignment, interaction guidance G_{inter} for physical correctness, and temporal smoothness guidance G_{temp} for motion consistency.

The silhouette guidance G_{sil} aligns the projected object geometry with the observed segmentation. We sample object surface points and project them onto the image plane to obtain $\mathcal{U}$. Let $\mathcal{B}$ denote sampled boundary pixels of the visible object mask, and let $\mathcal{M}$ be the union of the object mask and the dilated hand mask. The guidance is defined as

$$G_{\text{sil}} = \frac{1}{2|\mathcal{U}|} \sum_{\mathbf{u}\in\mathcal{U}} \rho(D_{\mathcal{M}}(\mathbf{u}) + D_{\Psi}(\mathbf{u})) + \frac{1}{2|\mathcal{B}|} \sum_{\mathbf{b}\in\mathcal{B}} \rho(D_{\mathcal{U}}(\mathbf{b})) , \tag{10}$$

where $D_{\mathcal{M}}(\mathbf{u})$ is the distance from $\mathbf{u}$ to the admissible region $\mathcal{M}$, $D_{\Psi}(\mathbf{u})$ projections outside the image, and $D_{\mathcal{U}}(\mathbf{b})$ is the distance from $\mathbf{b}$ to its nearest projected point in $\mathcal{U}$. Here, ρ denotes the smooth L1 penalty. The first term penalizes projections outside the object or hand-occluded regions, while the second term encourages coverage of the visible object boundary.

Our interaction guidance G_{inter} follows the interaction loss $\mathcal{L}_{\text{inter}}$ used during training, encouraging contact while penalizing interpenetration. Since ground-truth contact annotations are unavailable at test time, we use the learned contact prediction model HACO (Jung & Lee, 2026) to infer the contact index set Ω. This allows guidance to adapt to diverse hand-object interaction patterns without relying on manually predefined fixed contact regions (Fan et al., 2024; Hasson et al., 2019).

To further reduce temporal jitter, we introduce temporal smoothness guidance G_{temp} that penalizes the velocities and accelerations of the hand and object:

$$G_{\text{temp}} = \|\nabla \mathbf{s}^i\|_2^2 + \|\nabla^2 \mathbf{s}^i\|_2^2, \tag{11}$$

where ∇ and ∇^2 denote first- and second-order temporal finite differences, respectively. Translational increments are computed as differences between consecutive positions, while rotational increments are represented by the axis-angle vectors of the corresponding relative rotations. Second-order differences are computed from consecutive increments.

We apply guidance only during the later stages of flow-matching integration, when the predicted state is already close to a plausible interaction and requires only minor corrections. By concentrating refinement on these later steps, the guidance makes targeted adjustments that improve image alignment, interaction validity, and temporal consistency while introducing limited computational overhead.

Table 1: Results on HO3D-v3. Our method shows strong improvements on *interaction* metrics, which are often underreported in previous methods (Wang et al., 2025b; Shi et al., 2026). Hand and object accuracy are determined by foundation models and our method performs on par with AGILE (code not available).

Method	Interaction				Hand	Object			Robustness
	$CD_h \downarrow$	IV $\downarrow$	CA $\uparrow$	JM $\downarrow$	MPJPE $\downarrow$	CD $\downarrow$	F5 $\uparrow$	F10 $\uparrow$	SR $\uparrow$
HOLD	18.66	18.76	89.50	10.43	22.09	1.11	81.75	92.42	**100.0**
MagicHOI	21.81	17.41	83.00	21.66	7.38	0.90	76.74	91.59	83.3
AGILE	15.81	–	–	–	**3.92**	**0.27**	86.63	**97.77**	**100.0**
CHOIR	42.37	22.90	90.99	**0.92**	21.12	0.84	63.87	90.29	94.4
4D-HOF (ours)	**5.98**	**13.44**	**97.27**	1.61	4.40	0.28	**89.11**	97.13	**100.0**

3.4 Data Preparation and Model Training

We train our model on multiple large-scale HOI datasets, including HO-Cap (Wang et al., 2024), OakInk (Yang et al., 2022), and SHOWMe (Swamy et al., 2023), which covering diverse interactions and object categories. Training excludes all evaluation benchmarks, and we use identical model weights across test datasets to assess cross-dataset generalization.

To efficiently leverage the available data and prevent overfitting to specific dataset pattern, we construct training pairs by perturbing the ground-truth HOI states $\mathbf{s}_1$ with a randomly sampled noise (see Appendix for details). Unlike conventional generative modeling that transports samples from pure Gaussian noise, our formulation learns a conditional transport process between imperfect yet structured initialization states and physically plausible HOI states. This matches the practical inference setting, where modern foundation models already provide strong but imperfect estimates.

4 Experiments

4.1 Setup

Datasets. We evaluate our approach on two standard benchmarks: HO3D-v3 (Hampali et al., 2020) and DexYCB (Chao et al., 2021). HO3D-v3 features complex interactions resulting in severe occlusions, and DexYCB emphasizes tabletop grasping across diverse subjects. Following AGILE's evaluation protocol (Shi et al., 2026), we use 18 HO3D-v3 and 20 DexYCB sequences; we refer readers to AGILE for sequence selection and evaluation details.

Baselines. We compare with state-of-the-art monocular HOI reconstruction methods: HOLD (Fan et al., 2024) jointly optimizes implicit hand-object fields; MagicHOI (Wang et al., 2025b) improves object geometry with novel-view synthesis priors. AGILE (Shi et al., 2026) combines agentic initialization with heuristic optimization, and CHOIR (Xu et al., 2026) integrates foundation-model initialization, learned relative-depth correction, and contact-aware joint optimization.

Metrics. We evaluate five complementary aspects. *Hand accuracy*: mean per-joint position error (MPJPE) [mm] for local articulation. *Object accuracy*: bidirectional Chamfer distance (CD) [cm^2] and F-scores at 5/10 mm (F5/F10) [%] after ICP alignment of scale, rotation, and translation. *Interaction consistency*: hand-relative Chamfer distance (CD_h) [cm^2] (Fan et al., 2024), computed in the hand-root frame; interpenetration volume (IV) [cm^3], measuring hand-object overlap; and contact accuracy (CA) [%] (Jiang et al., 2021), the fraction of frames whose contact states match the ground truth. *Temporal stability*: hand-joint trajectory jerk magnitude (JM) [m/s^3], with lower values indicating smoother motion. *Robustness*: per-sequence success rate (SR) [%] (Shi et al., 2026), requiring both hand and object reconstruction to complete without failure.

4.2 4D HOI Reconstruction Comparison

We compare 4D-HOF against state-of-the-art methods on HO3D-v3 (Table 1) and DexYCB (Table 2), and show qualitative results in Figure 3 and the supplementary video.

Table 2: Results on DexYCB (Chao et al., 2021) dataset.

Method	Hand	Object			Interaction	Robustness
	MPJPE ↓	CD ↓	F5 ↑	F10 ↑	CD_h ↓	SR ↑
HOLD	30.86	19.30	33.20	54.94	170.90	45.0
MagicHOI	21.20	2.05	45.67	67.14	661.90	25.0
AGILE	19.06	0.52	83.21	95.43	94.60	**100.0**
CHOIR	27.24	0.84	79.46	93.09	134.74	**100.0**
4D-HOF (ours)	**6.84**	**0.45**	**83.62**	**95.81**	**48.85**	**100.0**

Table 3: Computational cost.

Method	Preproc.	Optim.	Total
4D-HOF	∼23 min	∼1 min	∼24 min
HOLD	∼65 min	∼24 h	∼25.1 h
MagicHOI	∼25 min	∼2 h	∼2.4 h
AGILE	21–29 min	∼1.75 h	2.1–2.2 h
CHOIR	∼11min	∼19min	∼30min

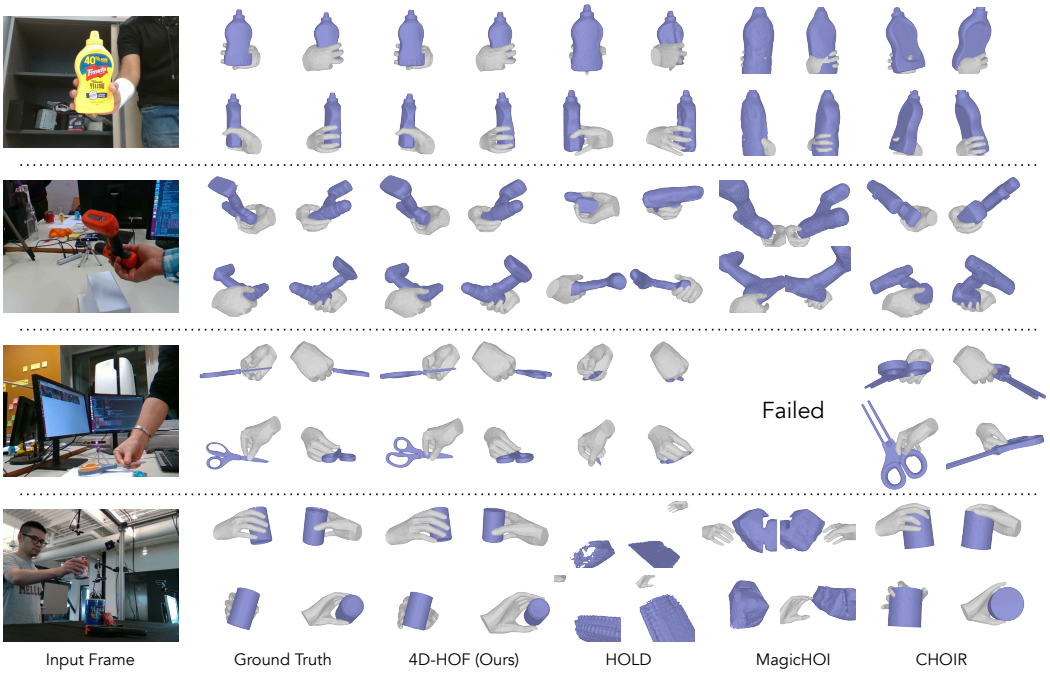

Figure 3: Qualitative results on HO3D-v3 (Hampali et al., 2020) and DexYCB (Chao et al., 2021). 4D-HOF consistently outperforms HOLD (Fan et al., 2024) and MagicHOI (Wang et al., 2025b) in preserving accurate geometry and hand-object interactions. Please see our supplementary video for detailed comparisons.

Results on HO3D-v3. 4D-HOF attains the lowest hand-to-object Chamfer distance and interpenetration volume among all methods, improving over the next best baseline by 62.2% in CD_h. The jerk magnitude is reduced to a level $6.4\times$ lower than HOLD (Fan et al., 2024) and $13.4\times$ lower than MagicHOI (Wang et al., 2025b), confirming that our learned interaction prior produces temporally stable trajectories without explicit smoothness regularization. Our method further achieves the highest contact ratio among reporting methods and the strongest F5 among all baselines.

Results on DexYCB. DexYCB contains diverse subjects performing tabletop grasps, where prior baselines suffer from large scale and translation errors. 4D-HOF yields a particularly pronounced margin across all the metrics comparing with other baselines.

Computational cost. We compare the inference time in Table 3. Existing methods rely on expensive test-time optimization and typically require hours to process a 200-frame sequence. In contrast, 4D-HOF performs reconstruction through a feed-forward generative transport process, reducing the refinement stage to approximately one minute while maintaining competitive reconstruction quality.

4.3 Ablation Study and Analysis

We conduct ablation studies on HO3D-v3 (Hampali et al., 2020) to validate the design choices.

How to use vision foundation models? We first study different strategies for constructing the initial hand-object states (Table 4). (a) Replacing our multi-view object generation with SAM3D (Chen et al., 2026a) significantly degrades interaction quality, showing the importance of temporally consistent

Table 4: Ablation of vision foundation models.

Method Variants	Interaction				Hand
	$CD_h \downarrow$	IV $\downarrow$	CA $\uparrow$	JM $\downarrow$	MPJPE $\downarrow$
4D-HOF (full model)	5.98	13.44	97.27	1.61	4.40
(a) SAM3D object recon.	262.19	9.87	64.40	1.75	6.09
(b) FPose default tracking	15646.78	9.39	75.74	2.07	5.00
(c) VFM only	46.58	11.06	78.07	3.95	4.03
(d) VFM + guidance	41.75	11.89	81.53	3.71	4.03

Table 5: Ablation of the bridge matching model.

Method Variants	Interaction				Hand
	$CD_h \downarrow$	IV $\downarrow$	CA $\uparrow$	JM $\downarrow$	MPJPE $\downarrow$
4D-HOF (full model)	5.98	13.44	97.27	1.61	4.40
(a) w/o temporal design	18.93	14.23	95.64	4.26	4.15
(b) w/o explicit encoding	11.92	14.76	96.41	1.62	4.40
(c) w/o object optimization	90.49	9.63	62.23	2.31	5.06
(d) w/o hand optimization	12.34	13.05	96.02	3.95	4.03
(e) w/o guidance	8.03	13.39	96.62	1.71	4.37

Table 6: Ablation studies on loss functions.

$\mathcal{L}$ Variants	Interaction				Hand
	$CD_h \downarrow$	IV $\downarrow$	CA $\uparrow$	JM $\downarrow$	MPJPE $\downarrow$
4D-HOF	5.98	13.44	97.27	1.61	4.40
w/o $\mathcal{L}_{3D}$	7.68	15.94	96.27	2.93	3.98
w/o $\mathcal{L}_{inter}$	6.02	13.78	97.00	1.63	4.60
w/o $\mathcal{L}_{reg}$	6.71	13.78	97.15	1.10	5.64

Table 7: Ablation studies on guidance terms.

G Variants	Interaction				Hand
	$CD_h \downarrow$	IV $\downarrow$	CA $\uparrow$	JM $\downarrow$	MPJPE $\downarrow$
4D-HOF	5.98	13.44	97.27	1.61	4.40
w/o G_{sil}	7.51	13.85	97.06	1.72	4.75
w/o G_{inter}	7.81	16.00	95.07	1.63	4.74
w/o G_{temp}	6.35	14.53	97.29	1.91	4.74

object reconstruction. (b) Using FoundationPose (Wen et al., 2024) in pure tracking mode also fails under severe occlusions, where a single mistracked frame can corrupt the entire sequence. (c) Directly using vision foundation model outputs without our generative bridge yields poor alignment and contact. (d) Adding guidance (but without bridge matching model) offers only modest improvements, suggesting the need for a learned interaction prior to refine independently estimated HOI states.

How to design the HOI flow matching model? We next analyze the design of the conditional bridge (Table 5). (a): Removing temporal modeling produces unstable 4D trajectories, suggesting that independent frame-wise refinement cannot capture interaction dynamics. (b): Removing the explicit hand-object structural encoding consistently degrades interaction metrics, highlighting the importance of explicitly modeling relative hand-object geometry. (c) and (d): Optimizing only the hand or only the object also leads to worse interaction quality, indicating that hand and object states must be refined jointly to resolve their mutual ambiguity. (e): Removing test-time guidance degrades contact and interaction quality, showing that guidance provides effective test-time control toward geometrically and physically consistent HOI states. Notably, these variants leave the object geometry unchanged, while contact quality and physical interaction degrade significantly.

How to supervise 4D-HOF training? We ablate each auxiliary losses in Table 6. Removing $\mathcal{L}_{3D}$ causes severe degradation across most metrics, showing that explicit 3D supervision is critical for HOI reconstruction. Removing $\mathcal{L}_{inter}$ reduces contact quality and hand pose accuracy, indicating that the contact loss encourages meaningful physical interaction. Removing $\mathcal{L}_{reg}$ leads to the largest drop in hand pose, suggesting that regularization helps preserve the hand state from foundation models.

How to guide 4D-HOF inference? We ablate each test-time guidance term in Table 7. Removing G_{sil} degrades all metrics, suggesting that image-space alignment helps refine the reconstructed interaction. Removing G_{inter} leads to the largest increase in interpenetration and the largest drop in contact accuracy, demonstrating its importance for plausible interactions. Removing G_{temp} causes the largest increase in motion jitter, confirming that temporal guidance improves motion consistency.

5 Conclusion and Future Works

We propose 4D-HOF, a novel feed-forward framework for 4D hand-object interaction reconstruction from monocular video. Our key innovation is to replace noise-driven generation with initialization derived from foundation models, and to learn a conditional generative bridge that transports these initial states toward physically plausible interactions. Our formulation enables joint refinement of hand, object, and their interactions in a feed-forward manner. Experiments show that our approach achieves strong performance in interaction quality while maintaining competitive hand and object accuracy, with significantly improved inference efficiency. One limitation, shared by many foundation model-based methods, is its sensitivity to initialization failures (Figure S6). Our current formulation focuses on single-hand interactions with rigid objects. Extending 4D-HOF to more complex settings, such as bimanual interactions and articulated or deformable objects, is an important future direction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding, literature search, and editing the paper to improve readability. We also used GPT-6 Astra in the agentic HOI reconstruction experiments reported in the Appendix. We reviewed and verified all AI-assisted work. LLM-generated code was tested by the authors, and all related works were individually reviewed and manually cited. We have disclosed all uses of generative AI that fall under the required disclosure tasks defined in the ICLR 2027 AI Policy for Authors. We take responsibility for the final content of this work, including text and code produced with the aid of generative AI.

ETHICS STATEMENT

Our work advances 4D hand-object interaction reconstruction from monocular video, with potential applications in robot learning, extended reality, and human-computer interaction. We do not collect new human-subject data, and our experiments use existing research datasets for training and evaluation under their respective terms of use.

Reconstructed hand motion and hand-object interactions from our method could be used to analyze how individuals manipulate objects in recorded videos, which may raise privacy concerns. Downstream applications should be deployed responsibly with appropriate consideration for privacy and permissions.

REPRODUCIBILITY STATEMENT

In this work, we propose 4D-HOF, a feed-forward framework for 4D hand-object interaction reconstruction from monocular video using conditional flow matching. We will release the code and trained model for our method upon acceptance. Further details regarding the initialization pipeline, model architecture, training procedure, implementation settings, and additional experiments can be found in the Appendix.

REFERENCES

Ahmed Tawfik Aboukhadra, Marcel Rogge, Nadia Robertini, Abdalla Arafa, Jameel Malik, Ahmed Elhayek, and Didier Stricker. Ghost: Fast category-agnostic hand-object interaction reconstruction from rgb videos using gaussian splatting. *arXiv preprint arXiv:2603.18912*, 2026.

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Ayce Idil Aytekin, Xu Chen, Zhengyang Shen, Thabo Beeler, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Grasp in gaussians: Fast monocular reconstruction of dynamic hand-object interactions. *arXiv preprint arXiv:2604.12929*, 2026a.

Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. In *Thirteenth International Conference on 3D Vision*, 2026b.

Blender Online Community. Blender: A 3d modelling and rendering package, 2026. URL https://www.blender.org.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

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

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026a.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. Forehoi: Feed-forward 3d object reconstruction from daily hand-object interaction videos. *arXiv preprint arXiv:2602.06226*, 2026b.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 494–504, 2024.

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

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11807–11816, 2019.

Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pp. 659–668. IEEE, 2021.

Jiawei He, Danshi Li, Xinqiang Yu, Zekun Qi, Wenyao Zhang, Jiayi Chen, Zhaoxiang Zhang, Zhizheng Zhang, Li Yi, and He Wang. Dexvlg: Dexterous vision-language-grasp model at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14248–14258, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Zehuan Huang, Yuan-Chen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16377–16387, 2025b.

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

Daniel Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. *Advances in Neural Information Processing Systems*, 38:120351–120384, 2026.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. Moge-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement. *arXiv e-prints*, pp. arXiv–2607, 2026.

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

Dixuan Lin, Tianyou Wang, Zhuoyang Pan, Yufu Wang, Lingjie Liu, and Kostas Daniilidis. Zero-shot reconstruction of in-scene object manipulation from video. *arXiv preprint arXiv:2512.19684*, 2025a.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022a.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022b.

Xingyu Liu, Pengfei Ren, Qi Qi, Haifeng Sun, Zirui Zhuang, Jing Wang, Jianxin Liao, and Jingyu Wang. Generalizable hand-object modeling from monocular rgb images via 3d gaussians. *Advances in Neural Information Processing Systems*, 38:127318–127340, 2025a.

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

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7037–7047, 2025b.

Haowen Luo, Yunze Liu, and Li Yi. Physics-aware hand-object interaction denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2341–2350, 2024.

Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015.

Kirill Mazur, Nikita Karaev, Matthew Chang, Jitendra Malik, and Nur Muhammad Shafiullah. AgentSTAR: Agentic shape tracking and reconstruction from monocular videos, 2026. URL https://arxiv.org/abs/2609.24487.

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17437–17447, 2025.

OpenAI. ChatGPT (6-astra) [large language model]. https://chatgpt.com, 2026. Accessed: 2026-09-25.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bhawna Paliwal, Haritheja Etukuru, William Liang, Pieter Abbeel, Nur Muhammad Mahi Shafiullah, and Jitendra Malik. Do as i do: Dexterous manipulation data from everyday human videos. *arXiv preprint arXiv:2606.19333*, 2026.

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

Vladimir I Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):677–695, 2002.

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

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

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

Jin-Chuan Shi, Binhong Ye, Tao Liu, Xiaoyang Liu, Yangjinhui Xu, Junzhe He, Zeju Li, Hao Chen, and Chunhua Shen. Agile: Hand-object interaction reconstruction from video via agentic generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. Showme: Benchmarking object-agnostic hand-object 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1935–1944, 2023.

Eugene Valassakis and Guillermo Garcia-Hernando. Handdgp: Camera-space hand mesh prediction with differentiable global positioning. In *European Conference on Computer Vision*, pp. 479–496. Springer, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

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

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

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. Arthoi: Taming foundation models for monocular 4d reconstruction of hand-articulated-object interactions. *arXiv preprint arXiv:2603.25791*, 2026.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–617. IEEE, 2023.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17868–17879, 2024.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5249–5260, 2025.

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7513–7524, 2026.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *International Conference on Learning Representations*, volume 2026, pp. 140700–140720, 2026.

Xiaofei Wu, Yi Zhang, Yumeng Liu, Yuexin Ma, Yujiao Shi, and Xuming He. Affordgrasp: Cross-modal diffusion for affordance-aware grasp synthesis. *arXiv preprint arXiv:2603.08021*, 2026.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4840–4851, 2024.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026.

Xianghui Xie, Bowen Wen, Chang Yan, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. CARI4D: Category agnostic 4d reconstruction of human-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2026.

Hao Xu, Yilin Liu, Yinqiao Wang, Chi-Wing Fu, and Niloy J Mitra. Choir: Contact-aware 4d hand-object interaction reconstruction. *arXiv preprint arXiv:2605.20992*, 2026.

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

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3895–3905, 2022.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19717–19728, 2023.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1920, 2024.

Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *arXiv preprint arXiv:2602.22209*, 2026.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

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

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2022.

Zhifan Zhu, Siddhant Bansal, Shashank Tripathi, and Dima Damen. Reconstructing objects along hand interaction timelines in egocentric video. *arXiv preprint arXiv:2512.07394*, 2025.

Table of Contents in Appendix

A Implementation Details **16**

A.1 Hand and Object State Initialization 16

A.1.1 Contextual scene parsing. 16

A.1.2 Hand reconstruction and state estimation. 16

A.1.3 Object reconstruction and state estimation. 16

A.2 4D HOI Flow Matching 17

A.2.1 Model architecture 17

A.2.2 Training data curation 18

A.2.3 Training details 18

B More Experimental Results **18**

B.1 HO3D-v3 MagicHOI's setting 18

B.2 Comparison of Object Generation 18

B.3 Do more HOI data help? 19

B.4 Comparison with GPT 6 Astra Max (as of Sep 20, 2026) 20

C Failure Case Analysis **22**

D VLM Prompt **22**

D.1 Interaction object reasoning 22

D.2 Multi-view image generation 22

D.3 Texture selection 23

A Implementation Details

In this section, we describe additional details of our method. We first present the hand and object state initialization pipeline based on various foundation models in Section A.1, followed by details of the flow-matching model in Section A.2.

A.1 Hand and Object State Initialization

A.1.1 Contextual scene parsing.

Given an input video $\mathcal{I}$, we first employ a VLM (Comanici et al., 2025) to understand the interaction and identify the object manipulated by the human hand. Since processing all video frames is computationally prohibitive, we uniformly sample five frames from the video and prompt the VLM to describe the target object. The resulting object description, together with the prompt "hand," is then provided to SAM 3 (Carion et al., 2026) for segmentation.

A.1.2 Hand reconstruction and state estimation.

For hand estimation, we use WiLoR (Potamias et al., 2025) to predict the MANO (Romero et al., 2022) parameters and 2D keypoints. We then recover the hand translation in camera coordinates using the camera intrinsics from MoGe-3 (Kong et al., 2026; Valassakis & Garcia-Hernando, 2024). Specifically, for each joint, the perspective projection yields the 2D position (u, v) of each 3D joint following equations:

$$u = f_x \frac{\mathbf{J}_x + \mathbf{t}_x}{\mathbf{J}_z + \mathbf{t}_z} + c_x, \tag{S1}$$

$$v = f_y \frac{\mathbf{J}_y + \mathbf{t}_y}{\mathbf{J}_z + \mathbf{t}_z} + c_y, \tag{S2}$$

where $(\mathbf{J}_x, \mathbf{J}_y, \mathbf{J}_z)$ denotes the position of 3D joint in the canonical hand space, $(\mathbf{t}_x, \mathbf{t}_y, \mathbf{t}_z)$ represents the global hand translation, and f_x, f_y, c_x, c_y are the camera intrinsic parameters. By aggregating constraints from all joints, we construct a linear system and solve it using least squares.

A.1.3 Object reconstruction and state estimation.

In order to automatically select a compact yet informative set of observations for the object from the video, we propose a visual feature–based frame selection strategy. Specifically, we extract image features using a pre-trained DINOv2 (Oquab et al., 2023) image encoder. Since DINO only yields patch-level embeddings, we also apply AnyUp (Wimmer et al., 2026) to enhance their spatial resolution. We then pool the upsampled features within the object mask region to obtain an object descriptor for each frame.

After having the object descriptors, the next step is picking the frames that are visually different. The frame with the largest object mask is first chosen as the anchor, and Farthest Point Sampling (FPS) is performed in the descriptor space to select a visual diverse set of frames. These selected frames are used as context prompts for canonical view generation, and the synthesized views are subsequently fed into a multi-view-to-3D model (Zhao et al., 2025) to recover the object mesh. Although the object geometry is generally reliable, textures produced by generative models are often unsatisfactory, preserving coarse color spots but failing on fine-grained patterns and text. To address this limitation, we construct two types of texture maps: one obtained directly from the generative model, which is suitable for unstructured objects with simple appearance, and another derived by projecting the canonical views (Huang et al., 2025b), which better preserves detailed patterns for structured objects. We render the object under multiple views using both texture types and employ a VLM to automatically select the optimal variant.

To better leverage FoundationPose (Wen et al., 2024) for object pose estimation, we initialize the pose on the frame with the largest object mask and propagate it bidirectionally throughout the sequence. Instead of relying on the default tracking mode of FoundationPose, in which a single occluded frame may lead to catastrophic error accumulation, we perform independent per-frame registration while explicitly rejecting pose hypotheses that deviate significantly from the previous estimate. Specifically,

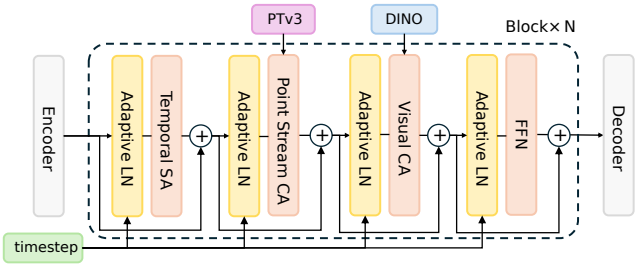

Figure S1: Model architecture. Our model adopts a DiT-style architecture for conditional flow matching. The latent hand-object state is processed through temporal self-attention, point-stream cross-attention using Point Transformer V3 features, and visual cross-attention using DINO features. Adaptive layer normalization conditioned on the timestep is applied throughout the network to modulate the generative transport process.

we use the geodesic distance between adjacent frame rotations as the filtering criterion and set the threshold to 60°.

We further employ silhouette-based filtering to further enforce visual consistency. For each pose hypothesis, we render an object mask $\tilde{\mathbf{M}}^o$ and compute its intersection-over-union (IoU) with the segmentation mask $\mathbf{M}^o$. To reduce the influence of hand occlusion, we exclude the hand mask $\mathbf{M}^h$ region before computing the IoU:

$$\text{IoU} = \frac{(\tilde{\mathbf{M}}^o \cap \sim \mathbf{M}^h) \cap (\tilde{\mathbf{M}}^o \cap \sim \mathbf{M}^h)}{(\tilde{\mathbf{M}}^o \cap \sim \mathbf{M}^h) \cup (\tilde{\mathbf{M}}^o \cap \sim \mathbf{M}^h)}. \tag{S3}$$

We retain only the top 30 pose hypotheses with the highest IoU scores. To further improve robustness, we enable both RGB and RGB-D modes in the pose refinement and scoring stages of FoundationPose, and compute the final pose score as the average of the two modalities.

A.2 4D HOI Flow Matching

A.2.1 Model architecture

Our flow matching model transports the noisy hand and object pose from the foundation model-based initialization to a plausible interaction pose manifold. Since we aim to reconstruct the 4D hand object interaction, a natural design is directly handling the sequence data, which benefits from the temporal information interaction. Specifically, our model takes a hand and object pose sequence as input. In each frame, the hand pose and object pose are concatenated together and tokenized by a linear layer. This pose token sequence is then passed thought several stacked Transformer blocks, after that the hand and object pose in each frame are decoded from the corresponding pose token.

There are three attention operations in each Transformer block. Our block starts with self-attention operation along the temporal dimension with RoPE (Su et al., 2024) for temporal embedding. This role of this layer is to utilize the temporal corrections in the sequence. In a motion sequence, the pose divergence usually small for nearby frames and large for far away frames, besides the motion evolution in the sequence should be smooth. By learning large scale data, this layer can make the pose sequences coherent and smooth. The second layer is a cross-attention layer between the 3D HOI point cloud sequence and pose token. The motivation of this operation is to inject explicit 3D spatial information to the model, since it's impossible to understand the hand object interaction solely from a set of bare rotations and translation. To prepare 3D cues, we use the pose results from the initialization stage to transform the MANO hand and object, then use a Point Transformer V3 (Wu et al., 2024) to encode them. In each frame, we use MANO vertices as hand point cloud and sample points on the object surface as the object point cloud. We treat the two components as a single point cloud and add a class label to each point. After feature extraction, a max pooling operation is applied to construct the final HOI point cloud token sequence. By this cross attention, the model can obtain explicit 3D information which boosts joint optimization. The third one is another cross attention with the 2D image features, since we hope the reconstruction results respect the observation. For each video frame, we use a frozen DINOv2-B to extract visual features and use Plucker ray to add camera information to each patch embedding. To balance the computational cost, we only apply

Table S1: Results on HO3D-v3 (Hampali et al., 2020) following MagicHOI's setting.

Method	Hand	Object				Interaction
	MPJPE ↓	CD ↓	F5 ↑	F10 ↑	RS ↓	CD_h ↓
iHOI	27.75	2.37	35.78	62.11	0.17	25.45
DiffHOI	16.02	2.30	39.59	64.49	0.13	33.33
EasyHOI	16.69	1.86	46.10	70.92	0.28	19.55
HOLD	30.79	1.31	57.20	80.23	0.62	21.28
MagicHOI	4.62	0.87	69.72	92.15	0.11	2.39
CHOIR	5.55	0.77	72.33	96.03	**0.10**	2.41
4D-HOF (ours)	**4.60**	**0.24**	**90.24**	**98.13**	0.13	**2.17**

cross attention within each frame. A standard feed-forward layer is added at the end of block, and adaptive layer norm is used before each attention and FFN later to inject flow-step information.

A.2.2 Training data curation

We initially explored training with initializations produced by our foundation model pipeline, but observed two limitations. First, object symmetries introduce ambiguities in the estimated rotations, complicating the construction of interpolation paths for bridge matching and subsequent training. Second, initialization errors exhibit dataset-specific biases, limiting the effectiveness of these initializations as a training source. We therefore adopt a dataset-agnostic perturbation strategy that synthetically corrupts ground-truth HOI states to emulate imperfect initializations.

We introduce synthetic noise to the ground-truth HOI states to emulate the imperfect initializations generated by the foundation models. To perturb the object pose, we sample a large offset (3.14 rad, 0.2 m) across 20% of the frames to simulate catastrophic pose estimation failures, while applying a smaller offset (0.1 rad, 0.02 m) to the remaining 80%. For the hand parameters, the translation vector is perturbed using noise sampled from a uniform spatial distribution (0.02, 0.02, 0.1). Meanwhile, the global orientation and the local pose of the hand are perturbed by adding Gaussian noise with a standard deviation of 0.1.

A.2.3 Training details

We optimize our model using AdamW with a learning rate of 0.0001 and a weight decay of 0.0001, regulated by a cosine learning rate scheduler with a warmup phase. The training process spans 20 epochs with a batch size of 16. During feature extraction, input images are resized to 224×224 before being encoded by DINOv2. To establish an explicit 3D encoding, we uniformly sample 4096 points from the object mesh surface. Finally, the temporal sequence length is set to 16 frames during training and extended to 128 frames during inference. The entire training process takes approximately 30 hours on a single NVIDIA H100 GPU.

B More Experimental Results

B.1 HO3D-v3 MagicHOI's setting

We empirically observe that some methods perform better on shorter clips, motivating an additional evaluation using the same sequences and frames as MagicHOI (Wang et al., 2025b). Table S1 and Figure S2 present the quantitative and qualitative results, respectively. Under this protocol, 4D-HOF achieves the best results on five of the six metrics, with substantial improvements in object reconstruction accuracy.

B.2 Comparison of Object Generation

We compare our multiview-based object generation pipeline with SAM3D (Chen et al., 2026a), with quantitative and qualitative results presented in Table S2 and Figure S3, respectively. Limited to

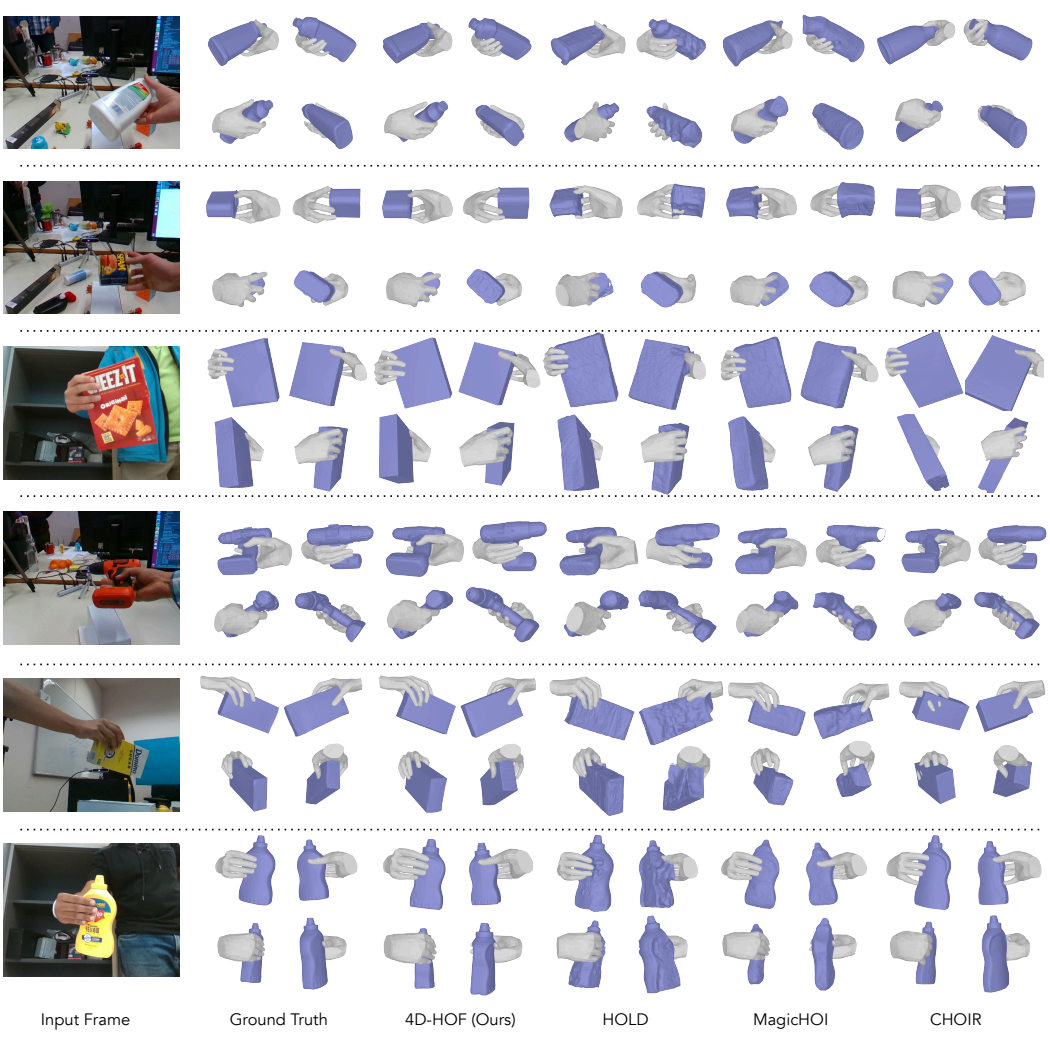

Figure S2: Qualitative results on HO3D-v3 (Hampali et al., 2020) following MagicHOI's setting.

Table S2: Object generation comparison. Our multiview-based pipeline outperforms SAM3D across all metrics, demonstrating the benefit of complementary visual observations for object reconstruction.

Methods	CD↓	F5↑	F10↑	RS↓
SAM3D	0.47	77.54	96.20	0.17
4D-HOF (Ours)	0.28	89.11	97.13	0.12

a single input image, SAM3D produces less reliable textures, which can compromise downstream pose estimation with FoundationPose, where rendered object appearance provides a key cue for pose evaluation. By combining complementary observations from multiple views with a powerful image generation model, our pipeline produces objects with higher geometric accuracy and texture fidelity.

B.3 Do more HOI data help?

To investigate the impact of training data volume, we evaluate our model's performance on the HO3D dataset using the CD_h metric (Figure S4). The results indicate that increasing the size of the training dataset yields consistent performance gains, demonstrating the scalability of our approach.



Figure S3: Qualitative comparison of object generation. We show two views of each object reconstructed by 4D-HOF and SAM3D (Chen et al., 2026a) alongside the ground truth. Our pipeline better preserves object geometry and texture details, including labels and fine structures. Ground-truth object widths are provided for scale reference.

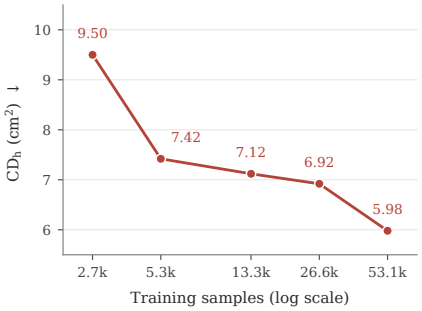

Figure S4: Dataset scaling analysis for bridge matching model training.

B.4 Comparison with GPT 6 Astra Max (as of Sep 20, 2026)

Recent GPT-6 (OpenAI, 2026) models have demonstrated strong capabilities in 3D reasoning and tool use. To provide a broader and more rigorous evaluation of our model, we constructed an agentic baseline by prompting GPT-6-Astra with Ultra reasoning effort and providing access to FoundationPose (Wen et al., 2024), WiLoR (Potamias et al., 2025), MoGe-3 (Kong et al., 2026), Blender (Blender Online Community, 2026), and related tools. We tasked the agent with reconstructing 4D hand–object interactions from the ABF12 and SM2 sequences of HO3D dataset , without imposing execution-time or token-budget limits. Our prompt is illustrated in B.4. Qualitative comparisons are shown in Figure S5. Although GPT-6 recovers plausible coarse hand and object poses, it exhibits noticeable hand–object separation and weaker cross-view contact consistency. In contrast, our method better preserves instance-specific object geometry and produces tighter, more physically plausible hand–object configurations: in SM2, the fingers more consistently wrap around the bottle, while in ABF12, our reconstruction better captures the complex fingertip contact and provide consistent result across viewpoints. These results suggest that, despite GPT-6's strong general-purpose 3D reasoning ability, the proposed pipeline provides more accurate geometry and more coherent hand–object interactions.

Beyond reconstruction quality, the agentic baseline incurs substantial runtime and resource overhead. In our experiments, a single run typically required several hours and consumed approximately 35% of the weekly usage allowance associated with our $100 subscription plan, making large-scale evaluation impractical under this budget. These observations highlight the challenges of applying general-purpose agentic systems to efficient, contact-consistent HOI reconstruction. Our feed-forward module could serve as a specialized component within such systems to accelerate reconstruction while improving geometric and interaction consistency.

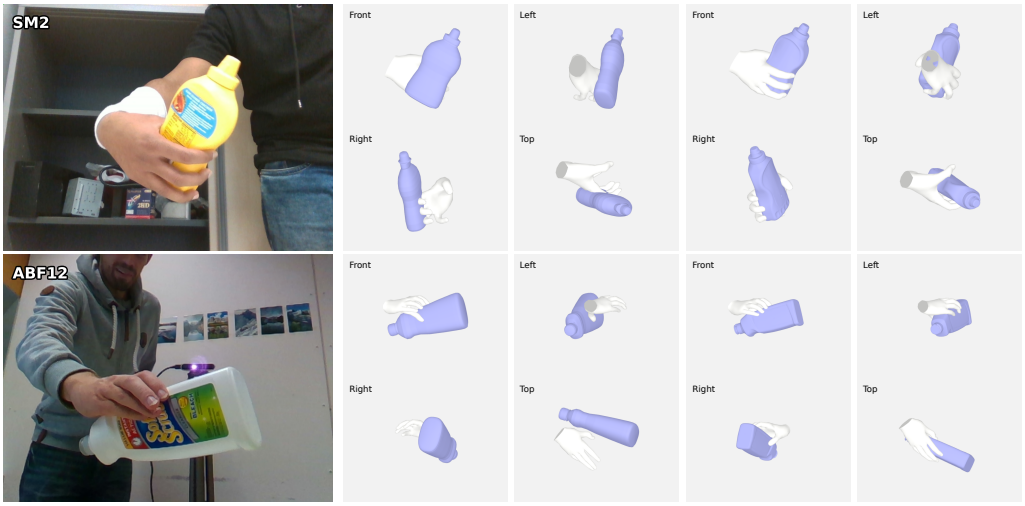

Figure S5: Qualitative comparison between GPT6-Astra and our 4D-HOF.

GPT6 HOI Reconstruction Prompt

Two third-person-view human hand–object interaction videos is available at: data/ABF12.mp4, data/SM2.mp4

For each video, reconstruct:
1. A temporally consistent MANO hand sequence.
2. A textured mesh for each manipulated object.
3. A per-frame 6DoF object pose sequence.

The hand and object should share a consistent metric coordinate system. Minimize hand–object penetration and separation, keep motion smooth, preserve plausible physical contact, and ensure rendered hand/object meshes align closely with the original RGB frames in 2D. Handle occlusion and tracking loss without large pose jumps, scale changes, or identity switches.

If additional Python packages are needed, create a virtual environment in the current directory rather than modifying the system environment. Blender and Blender-mcp are installed and enabled.

- Use MoGe-3 (https://github.com/microsoft/MoGe) for monocular depth estimation.
- Use WiLoR (https://github.com/rolpotamias/WiLoR) for MANO hand reconstruction, plus additional optimization procedures.
- Use FoundationPose (https://github.com/NVlabs/FoundationPose) for object 6DoF pose estimation, and make suitable modifications if you need. Create textured mesh and adjust mesh scale before running the model. You can use the depth from MoGe 3 as input.

Output:
- MANO parameters, joints, and vertices.
- Textured object meshes.
- Per-frame object poses and confidence.
- Camera parameters and coordinate-system documentation.
- RGB overlay videos and necessary 3D/diagnostic visualizations.
- README and other documents containing setup, commands, methods, optimization details, output paths, validation results, and known limitations.

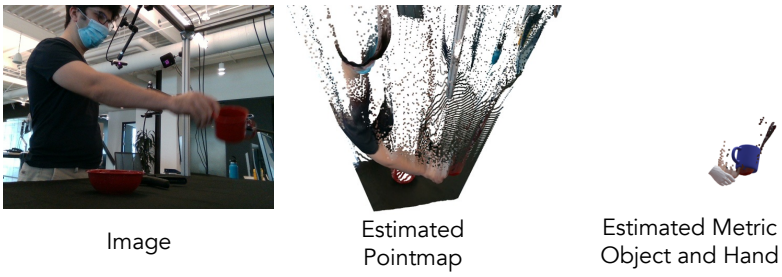

Figure S6: Failure analysis. When the estimated point map contains severe artifacts or out-of-distribution predictions, the resulting initialization may become unreliable, which can affect the flow-matching refinement process and lead to reconstruction failures.

C Failure Case Analysis

Our pipeline is reliant on several vision foundation models, with the monocular depth estimator, MoGe-3, acting as a critical geometric anchor. MoGe-3 (Kong et al., 2026) recovers the object's metric scale to match the hand and coarsely aligns both components within a unified coordinate space. Consequently, if the predicted point map exhibits significant artifacts (Figure S6), these inaccuracies cascade directly into the object scale and pose estimation modules. Such highly corrupted, out-of-distribution initializations ultimately misguide the flow-matching refinement model, resulting in catastrophic reconstruction failures.

D VLM Prompt

Our hand and object states initialization pipeline employs VLM for multiple purposes, including identifying the manipulated object, generating multi-view images, and selecting object textures. We describe each task below and provide the corresponding prompts.

D.1 Interaction object reasoning

Given an input video, the VLM identifies the object manipulated by the hand and generates a textual description. This description is then passed to SAM3 as a prompt for automatic object segmentation.

Object Identification Prompt

Describe the object interaction with hand, only return in [color object_name] format.

D.2 Multi-view image generation

We select object keyframes from the input video and provide them to an image generation model to synthesize canonical views of the object. The placeholder `{description}` in the prompt is filled with the object description obtained in the preceding reasoning step.

Canonical View Generation Prompt

Generate the front, back, left, right, top, bottom canonical views of this {description}. The generated views MUST perfectly match the exact shape, texture, color, and geometric details of the object in the reference images. Complete the occluded regions. Present as a 2*3 grid, front, back, left in the upper row, right, top, bottom in the lower row. Pure white background. The image must not contain text indicating viewpoint direction.

After each generation attempt, the VLM acts as a critic to assess whether the generated image satisfies the specified requirements. If the image fails this evaluation, we regenerate it.

Generated View Evaluation Prompt

You are an expert evaluator. Look at the provided generated image at the very end.
It is supposed to be a 2x3 image grid showing the front, back, left, right, top, and bottom canonical views of the object in the reference images.
Does the generated image strictly follow these requirements:
1. It is exactly a 2x3 grid.
2. Background is pure white.
3. No text indicating viewpoint direction.
4. Identity and texture are preserved.
Answer EXACTLY 'Yes' or 'No' without any explanation.

If both consecutive attempts fail, the VLM selects the better of the two candidates, allowing the pipeline to proceed.

Candidate Image Selection Prompt

I have two generated images at the end, each supposed to be a 2x3 grid showing canonical views of the same object from the preceding reference images. Image 1 is the previous best. Image 2 is the newly generated one.
Compare them based on these criteria:
1. Exactly a 2x3 grid layout.
2. Pure white background.
3. No text or labels.
4. High visual quality and accurate canonical views.
Which image strictly follows these instructions better and has better quality?
Return EXACTLY 'Image 1' or 'Image 2' without any explanation or markdown formatting.

D.3 TEXTURE SELECTION

Given the reconstructed object geometry, we obtain two candidate texture maps through direct generation and image projection, respectively. We apply each texture map to the geometry and render the resulting textured object from multiple viewpoints. The VLM then compares these multi-view renderings and selects the better texture.

Texture Selection Prompt

I have two 2x3 image grids showing canonical views of the same 3D object's geometry. Image 1 represents Model A. Image 2 represents Model B. Important: The 3D geometry is identical in both models, but the poses or viewpoints might differ between the two images.
Please act as a 3D texture quality judge. Ignore any pose or orientation differences, and focus strictly on comparing the texture quality. Look for detail, sharpness, realism, and a lack of noise or artifacts. Which model has better texture quality?
Return EXACTLY 'Model A' or 'Model B' without any explanation or markdown formatting.