EgoGrasp: World-Space Hand-Object Interaction Estimation from Egocentric Videos

Hongming Fu[1], Wenjia Wang[2†], Xiaozhen Qiao[3], Rolandos Alexandros Potamias[4], Taku Komura[2], Shuo Yang[5], Zheng Liu[6], and Bo Zhao[1‡]

[1] Shanghai Jiao Tong University [2] The University of Hong Kong
[3] University of Science and Technology of China [4] Imperial College London
[5] Harbin Institute of Technology (Shenzhen) [6] Beijing Academy of Artificial Intelligence
{fuhongming,bo.zhao}@sjtu.edu.cn wwj2022@connect.hku.hk
https://mint-sjtu.github.io/EgoGrasp.io

Abstract. We propose **EgoGrasp**, the first method to reconstruct world-space hand-object interactions (W-HOI) from dynamic egoview videos, supporting open-vocabulary objects. Accurate W-HOI reconstruction is critical for embodied intelligence yet remains challenging. Existing HOI methods are largely restricted to local camera coordinates or single frames, failing to capture global temporal dynamics. While some recent approaches attempt world-space hand estimation, they overlook object poses and HOI constraints. Moreover, previous HOI estimation methods either fail to handle open-set categories due to their reliance on object templates or employ differentiable rendering that requires per-instance optimization, resulting in prohibitive computational costs. Finally, frequent occlusions in egocentric videos severely degrade performance. To overcome these challenges, we propose a multi-stage framework: (i) a robust pre-processing pipeline leveraging vision foundation models for initial 3D scene, hand and object reconstruction; (ii) a body-guided diffusion model that incorporates explicit egocentric body priors for hand pose estimation; and (iii) an HOI-prior-informed diffusion model for hand-aware 6DoF pose infilling, ensuring physically plausible and temporally consistent W-HOI estimation. We experimentally demonstrate that EgoGrasp can achieve state-of-the-art performance in W-HOI reconstruction, handling multiple and open vocabulary objects robustly.

Keywords: Egocentric Video · Hand-Object Interaction · World Space

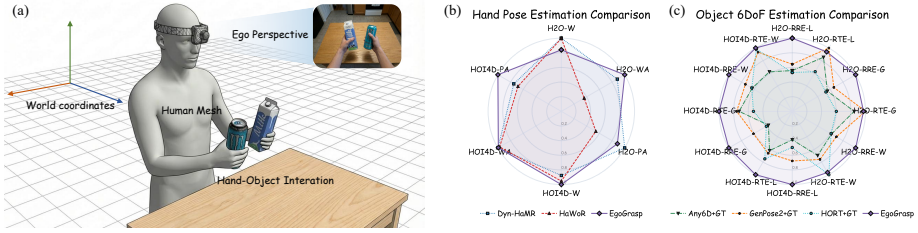

Fig. 1: (a) shows that EgoGrasp reconstructs world-space hand-object interactions from egocentric videos with dynamic cameras. (b) and (c) highlight the superior performance of EgoGrasp in estimating hand-object interactions in world space.

†: Project lead, ‡: Corresponding author

1 Introduction

Understanding hand-object interaction from egocentric videos is a cornerstone of computer vision and embodied intelligence. Reconstructing accurate world-space HOI poses is crucial for analyzing human manipulation behavior and enabling downstream applications in embodied AI, robotics, and virtual/augmented reality [12,20,38]. While egocentric videos provide unique first-person cues, they are typically recorded by dynamic cameras in unconstrained environments, where relying solely on per-frame camera-space geometry is insufficient. To fully interpret human actions, we argue that recovering temporally coherent trajectories in world space is essential to address three fundamental challenges: 1) Incorporating world-space is essential for stabilizing interaction dynamics. In egocentric scenarios, high-frequency camera jitter is entangled with object movement in local views. By anchoring the motion to the static world, we effectively filter out this noise, ensuring the reconstruction reflects firm physical grasps rather than unrealistic contact sliding. 2) Furthermore, static background cues could resolve semantic ambiguities. E.g., a user leaning towards an object is otherwise indistinguishable from lifting an object in the local frame. Global context clarifies the true source of motion, ensuring accurate trajectory reconstruction [34]. 3) World space helps ground the object within the static environment, enforcing geometric constraints that prevent objects from physically violating the scene, such as floating above or penetrating table surfaces during placement.

Despite progress in 3D hand and 6DoF object estimation, existing methods fail to predict world-space HOI in egocentric videos. Most approaches operate at the image or short-sequence level, predicting results in camera coordinates [3, 10, 18, 36, 39, 40], which change dynamically with the wearer's motion, making global trajectory recovery over time difficult. Considering object generalization, some rely on predefined CAD models [30,36]. Though some others [39,40] use differentiable rendering to generalize to different object categories, these methods are computationally expensive, sensitive to occlusions, and unstable in dynamic environments. Furthermore, current approaches underutilize egocentric priors, such as the structural coupling of the camera, body, and hands, limiting their robustness and generalization. As a result, reconstructing world-space HOI remains a major challenge, requiring methods that handle unknown objects, severe hand occlusions, and long-term spatial-temporal coherence without drift.

To address these challenges, we propose **EgoGrasp**, the first method, to the best of our knowledge, that reconstructs world-space hand–object interactions (W-HOI) from egocentric monocular videos with dynamic cameras. EgoGrasp adopts a multi-stage "perception–generation–infilling" framework that leverages reliable 3D cues from robust perception systems while introducing a generative motion prior to ensure temporal and global consistency. It operates in three stages: (1) Preprocessing: We recover accurate camera trajectories, scene geometry, and hand poses from egocentric videos with off-the-shelf SOTA models [16, 26]. We also leverage vision foundation models to reduce the complexity of the 3D problem to 2D, enabling the robust 6DoF initialization [1, 11, 32]. (2) Body Diffusion: To address the instability of egocentric perspectives and frequent

self-occlusions, we introduce a Body Diffusion model that synthesizes coherent body and hand motions by incorporating SMPL-X [25] upper-body constraints. To achieve superior accuracy, we integrate test-time optimization to enforce spatial alignment across the sequence. (3) HOI Diffusion: We leverage a diffusion model to infill discrete 6DoF sequences from stage one into plausible, continuous HOI trajectories based on the reconstructed hands from stage two. By modeling natural dynamics, it provides a high-quality initialization for final test-time optimization, resulting in precise and physically-grounded 6DoF trajectories.

We validate EgoGrasp on H2O and HOI4D datasets, achieving comparable results in world-space hand estimation and state-of-the-art results in world-space HOI reconstruction, with strong global trajectory consistency, demonstrating robustness to complex hand-object interactions and in-the-wild conditions.

Our key contributions are summarized as follows:

- We present a comprehensive analysis of the limitations inherent in current hand pose estimation, hand–object interaction modeling, and object 6DoF tracking approaches. Building upon these insights, we introduce the task of world-space hand–object interaction (W-HOI).
- We propose a novel framework, EgoGrasp, for W-HOI reconstruction from egocentric videos. EgoGrasp features a robust pre-processing stage and two generative prior diffusion models, enabling it to produce consistent HOI trajectories in world space. Furthermore, EgoGrasp is template-free and scalable to multiple objects, making it highly flexible and generalizable.
- Extensive experiments demonstrate that EgoGrasp substantially outperforms existing methods on the H2O and HOI4D datasets, thereby establishing new state-of-the-art results for W-HOI reconstruction in real-world settings, as shown in Fig. 1.

2 Related Work

Hand Pose Estimation. Hand Pose Estimation has developed rapidly in recent years, with early methods primarily targeting third-person perspectives under the assumption of minimal occlusion and stable camera viewpoints. Single-hand approaches typically regress MANO [28] model parameters [2], while two-hand methods employ implicit modeling or graph convolutions for interaction reconstruction [9,13]. Egocentric hand estimation is crucial for teaching robots manipulation tasks from a first-person perspective, facilitating advancements in embodied intelligence and virtual reality. Existing methods [17, 24, 27, 37] typically reconstruct hand poses in the camera coordinate system, limiting their ability to model hand-object interactions globally. To overcome this, recent studies have explored world-space pose estimation to recover hand poses and trajectories in world coordinates. For example, Dyn-HaMR [42] integrates SLAM-based camera tracking with hand motion regression to achieve 4D global motion reconstruction. Similarly, HaWoR [43] decouples hand motion from camera trajectories by leveraging adaptive SLAM and motion completion networks, enabling hand estimation in world coordinates. While recent approaches have improved hand

pose reconstruction, they fail to explicitly model the complex dynamics between hands and objects. Furthermore, current methods often underutilize egocentric priors, resulting in reduced robustness and generalization. To address these challenges, our EgoGrasp jointly models hand-object dynamics in world coordinates by incorporating with body-guided diffusion priors. Check Tab. 1 for differences between previous tasks. **Hand-Object Interaction Estimation.** Estimating hand pose and object 6DoF is inherently challenging, especially in hand-object interaction (HOI) scenarios, where the interactions between hand and object further increase the complexity. Existing object 6DoF estimation methods can be broadly categorized as: (1) template-based methods, which rely on predefined CAD models [30,36] and auxiliary inputs such as segmentation masks and depth maps; (2) template-free methods, which estimate the 6DoF pose without CAD models and may reconstruct the object mesh, often conditioned on RGB-D inputs and segmentation masks [10,15,35,44]. Typical approaches include combining a 3D generation model with 6DoF estimation method, or per-instance methods based on differentiable rendering or implicit surface learning. However, these approaches are often computationally expensive and struggle with robustness under noise, occlusions, and dynamic conditions.

HOI estimation builds on 6DoF methods by adding the challenge of estimating hand pose alongside object 6DoF. Template-based methods [3,5,18] only estimate hand pose and object 6DoF, while template-free methods [39,40] jointly reason about hand pose, object 6DoF, and object mesh reconstruction. Despite benefiting from joint reasoning, HOI methods face unique challenges such as severe occlusions, dynamic camera motion, and complex hand-object interactions. ContactOpt [5] and GraspTTA [7] both directly optimize the contact loss by predicting or generating hand-object contact heatmaps to better construct HOI results. DiffHOI [40] and G-HOP [39] also achieve object mesh reconstruction by leveraging differentiable rendering and an implicit SDF field guided by diffusion model priors. However, their computational cost is very high and their sensitivity to hand-object occlusion often results in poor temporal consistency and unstable motion reconstruction in real-world scenarios. To resolve the aforementioned issues, we present a robust 6DoF estimation pipeline based on a 3D foundation model SAM3D [32], an optical flow model MEMFOF [1] and a generative diffusion model for sequence refinement. This synergy enables seamlessly consistent 6DoF tracking and long-term stable mesh reconstruction. Tab. 1 highlights the differences between previous tasks.

Table 1: Comparison of representative tasks and world-space HOI. ✓: supported, ✗: not supported, –: partial/ambiguous.

Category	Ego	Hand Mesh	Obj 6DoF	Obj Mesh	World	Temp.
Exo Hand Est.	✗	✓	✗	✗	✗	–
Ego Hand Est.	✓	✓	✗	✗	✗	–
World Hand Est.	–	✓	✗	✗	✓	✓
Camera 6DoF	✗	✗	✓	–	✗	–
Camera HOI	✗	✓	✓	–	✗	–
W-HOI	✓	✓	✓	✓	✓	✓

Motion Prior Model In Pose Estimation All the aforementioned hand-only estimation methods suffer from a critical limitation: the excessive number of degrees of freedom. Due to this high dimensionality, these methods are highly sensitive to various noises, causing hand orientation and positional drift, depth

ambiguity, and even left–right hand misclassification. These issues fundamentally hinder stable world-space hand mesh reconstruction. Some motion prior models have been proposed to constrain the action representation within a reasonable range. VPoser [25] trains a pose prior network on large-scale MoCap data to constrain SMPL-X [25] parameters, aligning with human motion statistics. RoHM [45], LatentHOI [14], DiffHOI [40], and G-HOP [39] leverage diffusion models as priors for motion or HOI generation and reconstruction. Similarly, we construct decoupled diffusion prior models, including a body motion diffusion model and an HOI diffusion model, to learn the upper-body pose prior and HOI prior. The upper-body pose explicitly utilizes the egocentric prior, constraining hands by body that conform to the laws of motion.

3 Method

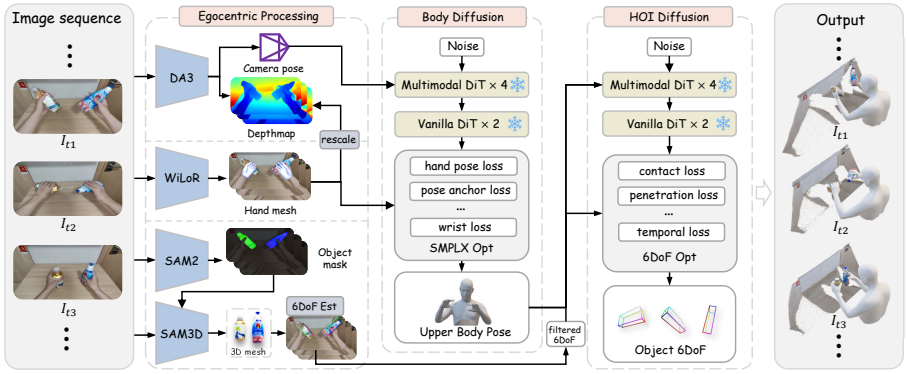

Fig. 2: Overview of EgoGrasp. We propose a three-stage method to recover world-space hand–object interaction from egocentric monocular videos with dynamic cameras: (1) extract 3D attributes with spatial perception models (for our 6DoF estimation method please see detail in Fig. 3); (2) reconstruct upper body pose with motion prior provided by Body Diffusion; (3) Interpolate discrete 6DoF sequences with HOI Diffusion for spatial, temporal, and contact consistency.

3.1 Problem Formulation

Given an egocentric video $V \in \mathbb{R}^{T\times H\times W\times 3}$, we aim to reconstruct accurate hand-object interactions in the world-space. Different from previous methods that treat left and right hand separately, we reconstruct the upper body pose to restrict the range of dual hands: hand poses $\{\theta_l^t, \theta_r^t \in \mathbb{R}^{15\times 3}\}_{t=0}^{T}$, body poses $\{\theta_b^t \in \mathbb{R}^{10\times 3}\}_{t=0}^{T}$ (upper-body only), betas $\beta \in \mathbb{R}^{10}$, global orientation and translation $\{\phi_t^i, \gamma_t^i \in \mathbb{R}^3\}_{t=0}^{T}$. For each object, we reconstruct the canonical mesh $\mathbf{M}$ and its global trajectory $\{o^t \in \mathrm{SE}(3)\}_{t=0}^{T}$ in world coordinates.

The proposed EgoGrasp method consists of three stages: 1) a preprocessing stage that extracts initial 3D hand poses and objects from the video using robust

off-the-self foundation models; 2) a body diffusion model that generates plausible upper body poses based on the extracted 3D attributes to constrain hand pose; and 3) an HOI diffusion model that infills the discrete 6DoF sequences from stage one into plausible and continuous sequences. An overview of the proposed method is visualized in Fig. 2.

3.2 Egocentric Video Preprocess

With the development of spatial intelligence, 3D tasks like point map reconstruction, parametric hand and body reconstruction, and 3D object reconstruction have also achieved significant progress, driven by advancements in foundation models and datasets. Based on this background, we leverage off-the-shelf foundational models to process egocentric videos in three steps: global scene reconstruction [16], hand reconstruction [26], and object reconstruction [32].

Step 1. Global Scene Reconstruction. For this step, we aim to get camera parameters and depth maps for the whole sequence. We utilize Depth-Anything3 (DA3) [16] to infer the camera intrinsics $\mathbf{K}$, extrinsics $\mathbf{E}^t$, represented as rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{T} \in \mathbb{R}^3$, along with depth maps D^t for the entire video sequence. The global scene reconstruction of this stage serves as the global coordinate system for the subsequent transformations

Step 2. Hand Reconstruction. We reconstruct MANO [28] parameters of both hands $\{\theta_l^t, \theta_r^t, \beta\}$ in the egoview videos by utilizing the state-of-the-art(SOTA) hand pose estimator WiLoR [26]. We then use the camera focal length obtained from the first step to calculate the MANO depth: $z = f/s$. The initially estimated depth map and camera translation from step 1 are aligned to the metric scale by multiplying a global scale factor. This factor is computed as the mean ratio between the depth values of the MANO model and the corresponding regions in the estimated depth map.

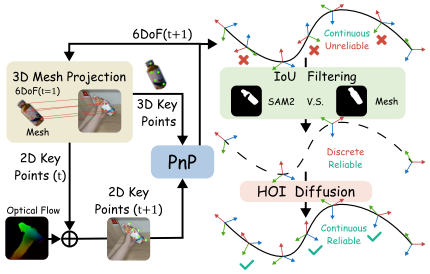

Fig. 3: Object 6DoF estimation pipeline. (1) We use PnP and optical flow to track 6DoF based on the initial SAM3D reconstruction on the first frame. (2) We filter out unreliable 6DoFs by mask IoU. (3) HOI Diffusion from EgoGrasp infills the reliable 6DoFs.

Step 3. Object Reconstruction. We propose a robust pipeline to reconstruct 6DoF poses of open-vocabulary objects, consisting of mesh initialization, temporal tracking, and trajectory refinement. First, to obtain a high-fidelity template mesh $\mathbf{M}$, we employ SAM3D [32] using its multi-view implementation [11] to enrich geometric details. This process yields the canonical shape $\mathbf{M}$ and the initial pose. To propagate this pose across the video, we utilize MEMFOF [1], an optical-flow-based tracker. By projecting 3D keypoints from $\mathbf{M}$ and tracking their 2D displacements, we solve the Perspective-n-Point (PnP) problem to recover a preliminary 6DoF trajectory. The initial tracking may suffer from drift, especially during hand-object occlusions. Instead

of relying solely on frame-wise tracking, we introduce a refinement stage. We first filter unreliable frames by checking the IoU between the SAM2 mask and the projected mesh mask, marking low-IoU frames as invalid (zero-padded). This sparse sequence $\mathbf{o}_{filtered}$ is then processed by our HOI Diffusion model. Leveraging generative priors, the model effectively infills the missing segments and rectifies the trajectory, producing physically plausible and continuous 6DoF motion (see Fig. 3). Further details of our HOI diffusion are in Sec. 3.4.

3.3 Body Diffusion Model

After obtaining the initial results from the preprocessing stage, we need a prior model with HOI knowledge to impose physical constraints and perform temporal completion. However, HOI datasets that include full-body poses are extremely limited. Training a unified model to simultaneously handle hand pose estimation and object 6DoF predictions under such constraints introduces substantial pose bias, which impedes the ability of the model to learn meaningful body motion priors. To address this, we propose to decouple the upper body pose estimation from the object 6DoF, training two separate networks, a Body Diffusion Model $\mathcal{W}$ and an HOI Diffusion Model $\mathcal{H}$, as shown in Fig. 2. We first utilize $\mathcal{W}$ to generate the upper-body pose, leveraging the body's limited reach range to constrain the hand's movement region, rather than treating both hands as independently translatable objects.

We additionally condition the model with features $\mathbf{c}^t$, extracted from a transformer condition encoder ϕ only from the central pupil frame (CPF) inter-frame transformations $\Delta\mathbf{T}_{\text{cpf}}^{t-1\to t}$. Following Egoallo [41], the CPF serves as a canonical reference coordinate system to bridge the SMPL-X local pose and the world space. Specifically, the origin of the CPF is anchored at the interocular midpoint of the SMPL-X model, while its axes are strictly aligned with the head orientation. The formulation of the body diffusion model's conditions is given as follows:

$$\Delta\mathbf{T}_{\text{cpf}}^{t\to t+1} = \left(\mathbf{E}^t\right)^{-1}\mathbf{E}^{t+1} \in \mathrm{SE}(3), \mathbf{c}^t = \phi\Big(\Delta\mathbf{T}_{\text{cpf}}^{t-1\to t}\Big). \tag{1}$$

Let $\mathbf{z}_0^{1:T} = \theta_{\text{body}}^{1:T}$ denote the ground truth SMPL-X body parameters. We formulate the body motion estimation as a conditional Denoising Diffusion Probabilistic Model (DDPM) [6]. Our body diffusion model $\mathcal{W}$ directly reconstructs the clean parameters $\mathbf{z}_0^{1:T}$ from the noisy state $\mathbf{z}_{t_d}^{1:T}$ at timestep t_d, conditioned on the features $\mathbf{c}^{1:T}$. The model is trained by minimizing the reconstruction error:

$$\mathcal{L}_{\text{W}} = \mathbb{E}_{t_d,\,\mathbf{z}_0,\,\boldsymbol{\epsilon}}\left[\left\|\mathcal{W}\big(\mathbf{z}_{t_d}^{1:T},\ \mathbf{c}^{1:T},\ t_d\big) - \mathbf{z}_0^{1:T}\right\|_2^2\right]. \tag{2}$$

During inference, the final predicted body parameters $\widehat{\theta}_{\text{body}}^{1:T}$ are obtained by denoising a sequence iteratively sampled from pure noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Test-time Optimization. At test time, we refine SMPL-X parameters with several objectives to ensure realistic and physically plausible body motion. (1) Pose

anchor loss $\mathcal{L}_{body}$ prevents excessive drift by keeping the optimized body configuration close to the predicted one, especially for arm and hand pose parameters. (2) Hand pose loss $\mathcal{L}_{pose}$ aligns the optimized left and right hand poses with the hand-pose predictions from WiLoR. (3) Hand 2D keypoint loss $\mathcal{L}_{kp2d}$ enforces image-space consistency by matching the projected 2D coordinates of our 3D hand joints to the corresponding WiLoR keypoints on jointly visible joints. (4) Wrist loss $\mathcal{L}_{wrist}$ enforces consistency of wrist 6DoF with WiLoR by constraining both wrist position and wrist orientation in the CPF coordinates. Please refer to the Supp.Mat for additional details.

3.4 HOI Diffusion Model

Our HOI diffusion model $\mathcal{H}$ is designed to generate plausible object 6DoF trajectories by leveraging reliable object predictions and human motion priors. It takes as input the initial reliable object 6DoF predictions $\mathbf{o}^t_{filtered}$, which are filtered by mask IoU and represented in wrist coordinates, alongside the conditions features $\mathbf{m}^t$ produced by a transformer condition encoder ϕ_{obj}. Specifically, the conditioning variables comprise: (1) the reliable object 6DoFs $\mathbf{o}^t_{filtered}$ in wrist coordinates; (2) a coarse translation $\mathbf{t}^t_{coarse}$ derived by projecting the masked depth using camera intrinsics and subsequently transformed into wrist coordinates; (3) a one-hot grasping label $\mathbf{l}^t_{grasp}$ determined by hand-to-object distances; (4) the hand poses θ^t_{obj} and hand joint locations $\mathbf{x}^t_{obj}$ in wrist coordinates generated by the body diffusion model $\mathcal{W}$. We train $\mathcal{H}$ using a masked reconstruction objective, to recover the randomly masked ground truth 6DoF trajectories. This enables effective training on existing HOI datasets while preserving the body motion priors learned from large-scale full-body datasets. The inference formula of HOI diffusion is as follows:

$$\mathbf{m}_j^{1:T} = \phi_{obj}\big(\mathbf{o}^{1:T}_{filtered}, \mathbf{t}^t_{coarse}, \mathbf{l}^{1:T}_{grasp}, \mathbf{x}^{1:T}_{obj}, \theta^{1:T}_{obj}\big), \tag{3}$$

$$\widehat{\mathbf{o}}^{1:T}_{t_d-1} = \mathcal{H}\big(\mathbf{o}^{1:T}_{t_d}, \mathbf{m}^{1:T}, \mathbf{o}^{1:T}_{filtered}, t_d\big), t_d \in [1, 1000], \tag{4}$$

where $\mathbf{o}$ denotes the 6DoF of each object. We train the HOI diffusion model $\mathcal{H}$ similar to the body diffusion model $\mathcal{W}$.

By looping through each object outside the HOI diffusion $\mathcal{H}$, multi-object interactions can be achieved, please check Supp.Mat. for details. Note that the object meshes $\mathbf{M}$ are obtained as described in 3.2.

Test-time Optimization. At test time, we perform a lightweight, fully differentiable optimization to refine object poses. The objective jointly enforces spatial accuracy and temporal smoothness for realistic and physically plausible object motion. (1) Object anchor loss $\mathcal{L}_{obj}$ prevents excessive drift by regularizing the refined object 6DoF toward the HOI diffusion prediction. (2) Contact loss $\mathcal{L}_{contact}$ enforces near-zero relative motion at contact by penalizing temporal displacement of grasping hand contact points in the object local frame, effectively reducing unnatural sliding. (3) Penetration loss $\mathcal{L}_{penetration}$ prevents non-physical interpenetration by penalizing hand points that violate a safety

margin to the object surface during complex HOI sequences. (4) Temporal loss $\mathcal{L}_{temporal}$ regularizes object motion over time by penalizing unstable velocity and acceleration in both rotation and translation, improving trajectory smoothness. For additional details please refer to the Supp.Mat.

4 Experiments

4.1 Implementation Details & Metrics

We evaluate hand pose estimation and object 6DoF estimation on two egocentric HOI datasets, the H2O [8] and HOI4D [19]. Following Dyn-HaMR [42] and HaWoR [43], the metrics employed for hand estimation evaluation included World Mean Per Joint Position Error (W-MPJPE, mm), World-aligned Mean Per Joint Position Position Error (WA-MPJPE, mm), and Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE, mm). For the 6DoF evaluation, we employ Relative Rotation Error (RRE, °) and Relative Translation Error (RTE, mm) across world, local, and wrist coordinate systems. Additionally, Chamfer Distance (CD, cm^2) in the wrist coordinates and Penetration Depth (PD, mm) are utilized to assess the geometric fidelity and physical plausibility of the HOI results. World-space metrics are computed over segments of 128 frames, where W-MPJPE involved aligning only the first two frames, whereas WA-MPJPE aligned the entire segment, both using Procrustes Alignment.

We train the model using PyTorch with 2 NVIDIA H20 GPUs at a learning rate of 2.5e-4, employing AdamW optimizer and cosine annealing. The body diffusion was trained on AMASS [22], 100STYLE [23], GRAB [31], PA-HOI [33], and HIMO [21] datasets; HOI diffusion was trained on GRAB [31], PA-HOI [33], and HIMO [21] datasets. Training sequences were sampled at 30 FPS with random lengths ranging from 64 to 256 frames. During test-time optimization, we used learning rates of 2.5e-4, 2.5e-4, and 1.0e-4 to optimize hand pose, body pose, and beta parameters, respectively, and performed a total of 50 optimization steps using AdamW and cosine annealing. For inference, we applied DDIM [29] sampling with 200 steps, and downsampled sequences by 3 at preprocessing stage.

4.2 World-Space Hand Pose Estimation

We demonstrate the superior reconstruction quality of EgoGrasp on the world-space hand pose by comparing it with the state-of-the-art Dyn-HaMR [42] and HaWoR [43] methods.

Quantitative Comparisons. Tab. 2 presents the quantitative results of EgoGrasp and other competing methods on the H2O and HOI4D datasets. As a complicated W-HOI framework, EgoGrasp achieves competitive results, reaching optimal or near-optimal performance compared to newly proposed SOTA hand-specific methods Dyn-HaMR and HaWoR. HaWoR performs the worst, with its metrics falling significantly behind both EgoGrasp and Dyn-HaMR. Furthermore, Dyn-HaMR face a performance drop (in PA-MPJPE) when processing long-range HOI trajectories within the HOI4D dataset.

Table 2: Hand pose evaluation on H2O and HOI4D datasets. For each metric, we use background colors to indicate the **best**, **second best**.

Method	H2O			HOI4D		
	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓
Dyn-HaMR (CVPR25)	**5.75**	46.37	**16.74**	9.83	**190.68**	63.12
HaWoR (CVPR25)	5.77	113.39	30.75	9.04	196.09	69.20
EgoGrasp	6.84	**40.93**	18.92	**8.61**	192.06	**47.29**

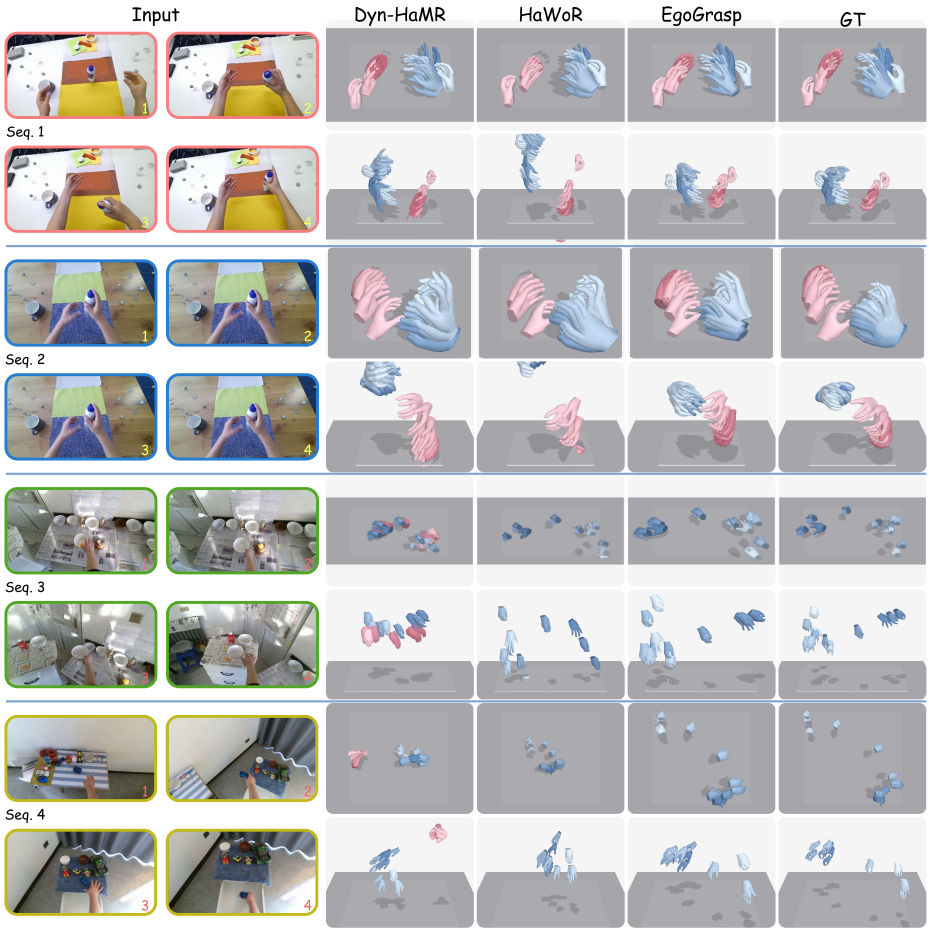

Fig. 4: World-space hand pose visualizations on the H2O dataset (top two sequences) and the HOI4D dataset (bottom two sequences).

Qualitative Comparisons. Fig. 4 shows a qualitative comparison of HaWoR, Dyn-HaMR, and EgoGrasp, demonstrating that EgoGrasp achieves superior performance in both fine-grained hand manipulation on the H2O dataset and long-term motion trajectories on the HOI4D dataset. As illustrated in the first two

sequences, EgoGrasp accurately reconstructs subtle hand movements, whereas the competing methods exhibit significant drift. Notably, HaWoR suffers from severe trajectory deviations, failing to produce a physically plausible path. In the subsequent sequences from HOI4D, HaWoR continues to show pronounced drifting and pose artifacts, while Dyn-HaMR even fails at hand-side identification (left vs. right). Furthermore, both baselines yield inaccurate world-space trajectories. In contrast, by incorporating a body pose constraint, EgoGrasp ensures kinematic consistency with natural human motion, resulting in substantially more precise and stable reconstructions.

4.3 World-Space Hand-Object Interaction Estimation

To further evaluate the effectiveness of EgoGrasp in the W-HOI task, we perform a comparative analysis against 3 baselines: 6DoF method Any6D [10] + WiLoR and GenPose2 [44] + WiLoR, as well as a feed-forward HOI model HORT [4]. We omit comparisons with per-instance optimization methods (e.g., differentiable rendering, implicit surface learning) due to their high time overhead and lack of scalability for large-scale HOI data. The 3 baselines differ in their output representations: HORT is a feed-forward model that only yields object point clouds based on WiLoR output, while GenPose2 is restricted to bounding box estimation and cannot generate meshes. Since HORT does not directly provide 6DoF poses, we employ the Iterative Closest Point (ICP) algorithm to register its predicted point clouds with the ground-truth meshes, thereby deriving the temporal 6DoF poses required for quantitative evaluation. It is important to note that the metrics in wrist coordinates are calculated by converting 6DoF or mesh to the corresponding hand wrist coordinates. For GenPose2, Any6D, and HORT, the hand results of WiLoR are used for calculating metrics in wrist coordinates.

Table 3: Object 6DoF estimation and HOI evaluation on H2O and HOI4D datasets. For each metric, we use background colors to indicate the **best** and **second best**.

Method	Local		Global		Wrist			
	RRE ↓	RTE ↓	RRE ↓	RTE ↓	RRE ↓	RTE ↓	CD ↓	PD ↓
Results on H2O Dataset								
GenPose2 (ECCV24)	28.88	**72.93**	28.53	**75.75**	33.19	74.19	-	-
Any6D (CVPR25)	33.40	85.43	33.77	89.62	37.27	80.82	52.51	3.18
HORT (ICCV25)	35.09	116.42	35.31	125.82	37.75	**56.22**	**22.98**	10.46
EgoGrasp	**18.62**	77.68	**18.71**	77.84	**23.08**	58.30	52.56	**0.74**
Results on HOI4D Dataset								
GenPose2 (ECCV24)	21.15	188.75	23.16	178.94	47.15	124.84	-	-
Any6D (CVPR25)	36.92	201.28	37.60	187.01	55.16	187.03	499.70	**0.46**
HORT (ICCV25)	29.04	166.36	34.67	256.26	54.63	118.90	**114.03**	5.56
EgoGrasp	**14.33**	**124.69**	**14.17**	**135.80**	**34.87**	**116.66**	235.53	0.64

Quantitative Comparisons. Tab. 3 presents the performance of various methods on the H2O and HOI4D datasets for object 6DoF tracking. Here, all other methods except EgoGrasp use ground-truth camera extrinsics to transform results from the camera coordinate system to the world coordinate system, whereas

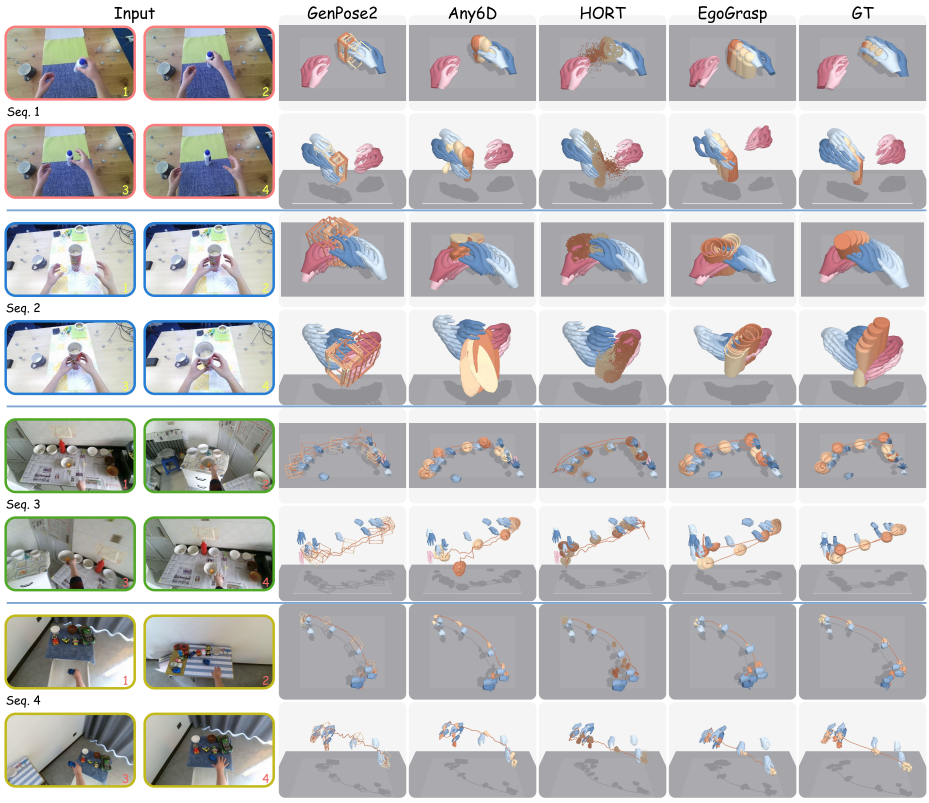

Fig. 5: World-space hand-object interaction visualizations on H2O (top two sequences) and HOI4D dataset (bottom two sequences).

EgoGrasp uses camera extrinsics predicted by DA3 and poses predicted by body diffusion to transform SMPL-X into world coordinates.

The experimental results demonstrate that EgoGrasp achieves substantial improvements in the RRE and RTE metrics across local, global, and wrist coordinate systems. In particular, the significant lead in the RRE metric underscores the robustness of EgoGrasp in modeling complex hand-object interactions, whereas other comparative methods remain limited to simple operations such as object translation. Although GenPose2 exhibits a marginal advantage in the RTE for local and global coordinates on the H2O dataset, the performance of this method suffers from severe degradation in the rotation and wrist coordinate systems. While HORT attains the leading performance in Chamfer Distance (CD), it functions as a single-frame method that only outputs point clouds, thus lacking temporal consistency and a complete mesh structure. This limitation leads to the highest Penetration Depth (PD) among all evaluated models, which compromises physical plausibility. In contrast, EgoGrasp maintains superior values in both the CD and PD metrics.

Table 4: Ablations study of hand pose estimation on H2O and HOI4D datasets. For each metric, we use background colors to indicate the **best**, **second**, and **worst**.

Method	H2O			HOI4D		
	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓
w/o $\mathcal{L}_{kp2d}$	9.60	47.25	22.06	8.97	**189.41**	48.32
w/o $\mathcal{L}_{pose}$	8.76	42.84	20.32	10.44	195.31	48.30
w/o $\mathcal{L}_{wrist}$	8.11	238.20	50.66	8.78	321.84	92.69
EgoGrasp	**6.84**	**40.93**	**18.92**	**8.61**	192.06	**47.29**

Qualitative Comparisons. As shown in Fig. 5, only EgoGrasp successfully reconstructs the object trajectory while simultaneously achieving accurate hand pose estimation. This superiority benefits from our body and HOI diffusion model, which jointly optimizes the initial hand pose and 6DoF estimations.

The initial two sequences demonstrate that GenPose2 is restricted to bounding box estimation, whereas the mesh generated by Any6D suffers from near-total failure, accompanied by substantial errors in the estimated 6DoF pose. Furthermore, the point cloud output produced by HORT fails to preserve the structural integrity of the object, manifesting significant divergence. In contrast, EgoGrasp achieves 6DoF estimation that remains fundamentally consistent with the ground truth while maintaining high-quality mesh reconstruction.

Regarding the subsequent two sequences, the three baseline methods exhibit varying degrees of trajectory jitter and physically implausible hand-object contact. EgoGrasp stands out by achieving robust reconstruction of long-range hand-object interactions within world-space coordinates.

4.4 Ablation Studies

To demonstrate the contribution of EgoGrasp component, we implemented several variants.

Ablations on Hand. Specifically, the variants "*w/o* $\mathcal{L}_{kp2d}$, *w/o* $\mathcal{L}_{pose}$, *w/o* $\mathcal{L}_{wrist}$" denote the removal of the corresponding loss terms during the SMPL-X test-time optimization process. Tab. 4 reports hand-pose ablation results on the H2O and HOI4D datasets. We observe that all variants except EgoGrasp exhibit significant declines in certain metrics, while EgoGrasp consistently achieves the best or competitive results. Notably, $\mathcal{L}_{wrist}$ demonstrates its important role in hand positioning, while $\mathcal{L}_{kp2d}$ and $\mathcal{L}_{pose}$ further optimize the poses based on it. The marginal decline of EgoGrasp in specific metrics (e.g., WA-MPJPE) compared to certain ablation variants is attributed to the synergistic interaction and trade-off between different loss terms, which prevents the optimization from over-fitting to a single metric. Overall, EgoGrasp demonstrates the most robust comprehensive performance.

Ablations on Object. "*w/o* $\mathcal{L}_{contact}$, *w/o* $\mathcal{L}_{temporal}$, *w/o* $\mathcal{L}_{penetration}$" refer to the exclusion of respective loss terms during object 6DoF test-time optimization. "*w/o* Opt." represents the raw results obtained directly from HOI diffusion without any post-optimization. Tab. 5 presents the object-6DoF ablation results on H2O and HOI4D. The performance trends mirror those observed in the

Table 5: Ablation study of object 6DoF estimation on H2O and HOI4D datasets. For each metric, we use background colors to indicate the **best**, **second**, and **worst**.

Method	Local		Global		Wrist			
	RRE ↓	RTE ↓	RRE ↓	RTE ↓	RRE ↓	RTE ↓	CD ↓	PD ↓
Results on H2O Dataset								
w/o $\mathcal{L}_{contact}$	**18.41**	105.67	**18.42**	107.10	**22.91**	77.27	54.89	0.98
w/o $\mathcal{L}_{temporal}$	19.12	78.81	19.19	79.27	23.45	**57.11**	**47.61**	0.93
w/o $\mathcal{L}_{penetration}$	18.88	78.85	18.95	79.28	23.28	57.20	**47.61**	0.93
w/o Opt.	19.48	102.57	19.49	104.16	24.20	68.98	49.16	0.81
EgoGrasp	18.62	**77.68**	18.71	**77.84**	23.08	58.30	52.56	**0.74**
Results on HOI4D Dataset								
w/o $\mathcal{L}_{contact}$	15.22	144.85	15.22	155.15	34.99	132.88	251.34	0.87
w/o $\mathcal{L}_{temporal}$	15.37	126.75	15.25	137.32	**34.81**	117.26	229.23	0.88
w/o $\mathcal{L}_{penetration}$	15.27	126.71	15.15	137.23	34.85	117.28	**228.77**	0.88
w/o Opt.	15.25	142.32	15.18	153.36	35.00	128.39	245.52	0.78
EgoGrasp	**14.33**	**124.69**	**14.17**	**135.80**	34.87	**116.66**	235.53	**0.64**

hand-pose ablations: most variants show a substantial drop in specific indicators. For instance, removing $\mathcal{L}_{contact}$ leads to a severe decline in RTE and CD, highlighting its necessity for accurate translation. The absence of $\mathcal{L}_{temporal}$ results in a significant increase in RRE, underscoring its importance for rotational consistency. Furthermore, $\mathcal{L}_{penetration}$ effectively minimizes physical violations, as evidenced by the degraded PD metric when it is removed. While the "*w/o* Opt." yields acceptable results, which validates the strong infilling capacity of our HOI diffusion, it still underperforms across all metrics compared to the full EgoGrasp, proving the necessity of test-time optimization. Although EgoGrasp slightly trails in isolated metrics due to the balancing effect of multiple constraints, it remains best in overall metrics.

5 Conclusion

We introduced EgoGrasp, the first method to reconstruct world-space hand–object interactions (W-HOI) from egocentric monocular videos captured by dynamic cameras. Our multi-stage framework integrates a robust pre-processing pipeline built on vision foundation models, an body-guided prior diffusion model for hand estimation, and a template-free object 6DoF estimation pipline based on an HOI diffusion model. EgoGrasp yields accurate, physically plausible, and temporally coherent W-HOI trajectories that generalize beyond single-object and template constraints. Experiments on H2O and HOI4D datasets demonstrate its advanced performance under long-range motion and severe hand-object occlusion. Looking ahead, we plan to explore more streamlined feed-forward architectures for modeling world-space human–object interaction, with the goal of reducing reliance on preprocessing while preserving accuracy and robustness.

References

1. Bargatin, V., Chistov, E., Yakovenko, A., Vatolin, D.: Memfof: High-resolution training for memory-efficient multi-frame optical flow estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8187–8196 (2025)
2. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10843–10852 (2019)
3. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12417–12426 (2021)
4. Chen, Z., Potamias, R.A., Chen, S., Schmid, C.: Hort: Monocular hand-held objects reconstruction with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6046–6057 (2025)
5. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1471–1481 (2021)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
7. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11107–11116 (2021)
8. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10138–10148 (October 2021)
9. Lee, J., Sung, M., Choi, H., Kim, T.K.: Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21169–21178 (2023)
10. Lee, T., Wen, B., Kang, M., Kang, G., Kweon, I.S., Yoon, K.J.: Any6d: Model-free 6d pose estimation of novel objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11633–11643 (2025)
11. Li, B.: Mv-sam3d: Adaptive multi-view 3d reconstruction. `https://github.com/devinli123/MV-SAM3D` (2025), gitHub repository, commit 65b6a3e, accessed: 2026-03-04
12. Li, K., Li, P., Liu, T., Li, Y., Huang, S.: Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6991–7003 (2025)
13. Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T., Liu, Y.: Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2761–2770 (2022)
14. Li, M., Christen, S., Wan, C., Cai, Y., Liao, R., Sigal, L., Ma, S.: 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)
15. Li, Z., Wang, G., Ji, X.: Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7678–7687 (2019)

16. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
17. Liu, R., Ohkawa, T., Zhang, M., Sato, Y.: Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 677–686 (2024)
18. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3d hand-object poses estimation with interactions in time. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14687–14697 (2021)
19. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21013–21022 (June 2022)
20. Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., Lu, Z.: Being-h0: vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597 (2025)
21. Lv, X., Xu, L., Yan, Y., Jin, X., Xu, C., Wu, S., Liu, Y., Li, L., Bi, M., Zeng, W., et al.: Himo: A new benchmark for full-body human interacting with multiple objects. In: European Conference on Computer Vision. pp. 300–318. Springer (2024)
22. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
23. Mason, I., Starke, S., Komura, T.: Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. Proceedings of the ACM on Computer Graphics and Interactive Techniques **5**(1), 1–18 (2022)
24. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12999–13008 (2023)
25. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
26. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: 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)
27. Prakash, A., Tu, R., Chang, M., Gupta, S.: 3d hand pose estimation in everyday egocentric images. In: European Conference on Computer Vision. pp. 183–202. Springer (2024)
28. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
30. Su, Y., Saleh, M., Fetzer, T., Rambach, J., Navab, N., Busam, B., Stricker, D., Tombari, F.: Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6738–6748 (2022)

31. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020), https://grab.is.tue.mpg.de
32. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025), https://arxiv.org/abs/2511.16624
33. Wang, R., Zuo, L., Lin, Z., Wang, Q., Cheng, Z., Xie, R., Ling, J., Song, L.: Pa-hoi: A physics-aware human and object interaction dataset. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 12769–12775 (2025)
34. Wang, W., Pan, L., Pi, H., Lou, Y., Ren, X., Wu, Y., Liao, Z., Yang, L., Dabral, R., Theobalt, C., et al.: Embodmocap: In-the-wild 4d human-scene reconstruction for embodied agents. arXiv preprint arXiv:2602.23205 (2026)
35. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–617 (2023)
36. Wen, B., Yang, W., Kautz, J., Birchfield, S.: 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)
37. Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., Wang, W.: Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21243–21253 (2023)
38. Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.C., Zou, X., Fang, Y., Cheng, X., Qiu, R.Z., et al.: Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440 (2025)
39. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: 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)
40. Ye, Y., Hebbar, P., Gupta, A., Tulsiani, S.: Diffusion-guided reconstruction of everyday hand-object interaction clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19717–19728 (2023)
41. Yi, B., Ye, V., Zheng, M., Li, Y., Müller, L., Pavlakos, G., Ma, Y., Malik, J., Kanazawa, A.: Estimating body and hand motion in an ego-sensed world. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7072–7084 (2025)
42. Yu, Z., Zafeiriou, S., Birdal, T.: Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27716–27726 (2025)
43. Zhang, J., Deng, J., Ma, C., Potamias, R.A.: Hawor: World-space hand motion reconstruction from egocentric videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1805–1815 (2025)
44. Zhang, J., Huang, W., Peng, B., Wu, M., Hu, F., Chen, Z., Zhao, B., Dong, H.: Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In: European Conference on Computer Vision. pp. 199–216. Springer (2024)
45. Zhang, S., Bhatnagar, B.L., Xu, Y., Winkler, A., Kadlecek, P., Tang, S., Bogo, F.: Rohm: Robust human motion reconstruction via diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14606–14617 (2024)

A Transformations

Fig. 6 illustrates the upper-body SMPL-X mesh generated by EgoGrasp, together with the visualizations of three coordinate frames: LeftWrist, RightWrist, and CPF. The rigid transformations among these coordinate frames are used as conditions and inputs for the body diffusion model and SMPL-X test-time optimization, guiding EgoGrasp to synthesize upper-body poses that are consistent with the egocentric viewpoint priors. It also shows the object coordinates, note that HOI Diffusion infills object 6DoF in wrist coordinates, then it can be transformed into CPF coordinates.

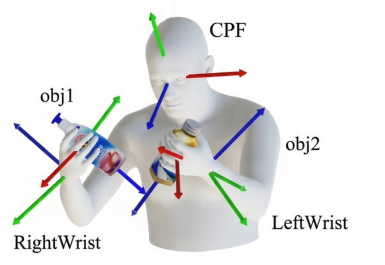

Fig. 6: Transformations visualization. The coordinate axes in the figure represent the corresponding coordinate system transformations.

B 6DoF Tracking & HOI Diffusion Infilling

This note describes Step 3 (*Object Reconstruction*) of the main paper. Let the video contain frames $t = 1, \ldots, T$, with RGB image I_t, depth map D_t, camera intrinsics $K_t \in \mathbb{R}^{3\times3}$, and SAM2 object mask $S_t \subset \Omega$, where Ω denotes the image domain. And the object pose in camera coordinates is

$$T_t^{o\to c} = \begin{bmatrix} R_t & \mathbf{t}_t \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3), \tag{5}$$

where $R_t \in SO(3)$ and $\mathbf{t}_t \in \mathbb{R}^3$.

Anchor-frame reconstruction. An anchor frame a is selected and provides an image I_t, a mask S_t, and a masked pointmap $\mathbf{P}_t$. These observations are fed to SAM3D, which outputs a canonical mesh $\mathbf{M}$, an anchor-frame pose, and a mesh scale $\mathbf{s}$. And we use the scaled mesh $\mathbf{M_s}$.

Pose propagation by MEMFOF and PnP. Starting from the anchor pose, 6DoF tracking is performed on sampled frames and then interpolated to the full video. Let $F_{t\to t+1} : \Omega \to \mathbb{R}^2$ denote the optical flow from frame t to frame $t+1$ predicted by MEMFOF. Given the current pose $T_t^{o\to c}$, surface points $\{\mathbf{X}_i\}$ are sampled on $\mathbf{M_s}$, projected to frame t, and kept only if they fall inside S_t. Then the optical flow transports each visible point to the next frame.

$$\mathbf{x}_i^t = R_t\mathbf{X}_i + \mathbf{t}_t, \qquad \mathbf{u}_i^t = \pi_{K_t}(\mathbf{x}_i^t), \qquad \tilde{\mathbf{u}}_i^{t+1} = \mathbf{u}_i^t + F_{t\to t+1}(\mathbf{u}_i^t). \tag{6}$$

where $\pi_{K_t}(\cdot)$ denotes standard perspective projection under intrinsics K_t. This gives candidate 2D–3D correspondences:

$$\mathcal{C}_{t\to t+1} = \left\{ \left(\mathbf{X}_i, \tilde{\mathbf{u}}_i^{t+1}\right) \right\}. \tag{7}$$

The next pose is then estimated by RANSAC-PnP followed by refinement:

$$T_{t+1}^{o\to c} = \arg\min_{T\in SE(3)} \sum_{i\in\mathcal{I}_{t+1}} \left\|\pi_{K_{t+1}}(T\mathbf{X}_i) - \tilde{\mathbf{u}}_i^{t+1}\right\|_2^2, \tag{8}$$

where $\mathcal{I}_{t+1}$ indexes the 2D–3D correspondences. Propagating forward and backward from the anchor yields all the frames.

Mask-IoU reliability. For each frame, the current mesh is rasterized to a binary mask. And hand-occluded pixels are removed by hand mask H_t.

$$\widehat{S}_t = \mathcal{R}(\mathbf{M_s}, T_t^{o\to c}, K_t), \qquad \widehat{S}_t^{\text{vis}} = \widehat{S}_t \setminus H_t, \tag{9}$$

where $\mathcal{R}(\cdot)$ denotes triangle rasterization. And the reliability score is the mask IoU between S_t and $\widehat{S}_t^{\text{vis}}$. If the pose is unavailable or the object mask is empty, we set $\gamma_t = 0$. Given a threshold τ_{IoU}, the trusted-frame indicator is

$$m_t = \mathbb{1}[\gamma_t \geq \tau_{\text{IoU}}]\,. \tag{10}$$

Let $(\mathbf{r}_t^{\text{raw}}, \mathbf{u}_t^{\text{raw}})$ denote the raw pose in the diffusion representation, where $\mathbf{r}_t$ is the 6D rotation representation and $\mathbf{u}_t$ is the translation. The filtered 6DoFs is

$$\mathbf{r}_t^{\text{filtered}} = \begin{cases} \mathbf{r}_t^{\text{raw}}, & m_t = 1, \\ \mathbf{0}, & m_t = 0, \end{cases} \qquad \mathbf{u}_t^{\text{filtered}} = \begin{cases} \mathbf{u}_t^{\text{raw}}, & m_t = 1, \\ \mathbf{0}, & m_t = 0. \end{cases} \tag{11}$$

Interaction with HOI diffusion. The HOI diffusion model receives both the motion hypothesis and the IoU-filtered trusted branch. Let $g_t \in \{L, R, \varnothing\}$ denote the grasp label computed roughly by hand-object distance, where $g_t = \varnothing$ means that no grasp is active. The trusted 6DoF condition $\widetilde{\mathbf{o}}_t^{\text{cond}}$ is provided only when both trust and grasp validity hold, otherwise padded with 0. During HOI Diffusion sampling, it predicts trajectory only on untrusted grasp frames:

$$\Omega_{\text{pred}} = \{t \mid g_t \neq \varnothing,\ m_t = 0\}\,. \tag{12}$$

Thus, the IoU filter determines which frames act as hard anchors and which are corrected or completed by the HOI prior.

$$\begin{aligned} \text{SAM3D} &\longrightarrow \text{MEMFOF + PnP} \longrightarrow \text{interpolation} \\ &\longrightarrow \text{mask-IoU filtering} \longrightarrow \text{HOI diffusion} \end{aligned} \tag{13}$$

The key design is to keep temporal propagation conservative and geometry-driven, and to invoke the generative prior only where the trajectory is marked as untrusted and grasped.

Summary. Algorithm 1 summarizes the full Step-3 pipeline. It first reconstructs a canonical mesh and anchor pose by SAM3D and applies the predicted mesh scale. It then propagates poses over sampled frames using MEMFOF-induced correspondences and RANSAC-PnP. Next, the posed mesh is rasterized in each frame to compute a hand-aware mask IoU reliability score, which determines

Algorithm 1 SAM3D–MEMFOF Tracking with HOI Diffusion Infilling

Require: $\{I_t, D_t, K_t, S_t, \mathbf{P}_t, H_t, g_t\}_{t=1}^{T}$, τ_{IoU}, HOI context $\mathbf{c}_{\text{HOI}}^{1:T}$
Ensure: Raw poses $\{T_t\}_{t=1}^{T}$, trusted branch $\{\mathbf{o}_t^{\text{trusted}}\}_{t=1}^{T}$, completed poses $\{\widehat{\mathbf{o}}_t\}_{t=1}^{T}$

Select anchor a, support views $\mathcal{V}_a$
$(\mathbf{M_s}, T_a) \leftarrow \mathsf{SAM3D}(\{I_t, S_t, \mathbf{P}_t\})$
for each sampled edge (t, t') in the forward/backward pass from a **do**
  $F_{t\to t'} \leftarrow \mathsf{MEMFOF}(t \to t')$
  $\mathcal{C}_{t\to t'} \leftarrow \mathsf{FlowCorr}(\mathbf{M_s}, T_t, K_t, S_t, F_{t\to t'}, S_{t'}, H_{t'})$
  $T_{t'} \leftarrow \mathsf{PnP}(\mathcal{C}_{t\to t'}, K_{t'})$
end for
for $t = 1$ to T **do**
  $\widehat{S}_t^{\text{vis}} \leftarrow \mathsf{RenderMask}(\mathbf{M_s}, T_t, K_t) \setminus H_t$
  $\gamma_t \leftarrow \mathsf{IoU}(S_t, \widehat{S}_t^{\text{vis}}), \quad m_t \leftarrow \mathbb{1}[\gamma_t \geq \tau_{\text{IoU}}]$
  $\mathbf{o}_t^{\text{filtered}} \leftarrow m_t\, \mathbf{o}_t^{\text{raw}}, \quad \widetilde{\mathbf{o}}_t^{\text{cond}} \leftarrow \mathbb{1}[m_t = 1 \wedge g_t \neq \varnothing]\, \mathbf{o}_t^{\text{filtered}}$
end for
$\Omega_{\text{pred}} \leftarrow \{\, t \mid g_t \neq \varnothing,\ m_t = 0 \,\}$
$\{\widehat{\mathbf{o}}_t\}_{t=1}^{T} \leftarrow \mathsf{HOIDiffusion}(\{\mathbf{o}_t^{\text{filtered}}\}_{t=1}^{T}, \{\widetilde{\mathbf{o}}_t^{\text{cond}}\}_{t=1}^{T}, \mathbf{c}_{\text{HOI}}^{1:T}, \Omega_{\text{pred}})$
return $\{T_t\}_{t=1}^{T}$, $\{\mathbf{o}_t^{\text{filtered}}\}_{t=1}^{T}$, $\{\widehat{\mathbf{o}}_t\}_{t=1}^{T}$

the trusted-frame indicator. The raw pose is converted to the diffusion representation, from which the filtered 6DoFs and the grasp-gated conditioning signal are constructed. Finally, HOI diffusion predicts only on untrusted grasp frames, while trusted frames remain fixed as anchors.

Here, $\mathsf{FlowCorr}(\cdot)$ denotes flow-based 2D–3D correspondence construction, $\mathsf{RenderMask}(\cdot)$ denotes mesh rasterization, $\mathsf{IoU}(\cdot)$ denotes reliability computation, $\mathsf{HOIDiffusion}(\cdot)$ denotes untrusted 6DoF infilling.

C Multi-Object Inference Loop

Fig. 7 and Algorithm 2 describe our decoupled multi-object human–object interaction estimation pipeline. The goal is to recover a temporally coherent upper-body motion together with physically plausible 6DoFs for all interacted objects over a time window of length T. The method is decomposed into two stages: the body diffusion model $\mathcal{W}$ first infers human motion, and the HOI diffusion model $\mathcal{H}$ then refines each object trajectory conditioned on the recovered human motion. This decomposition avoids coupling body-pose generation and object-pose generation in a single model while

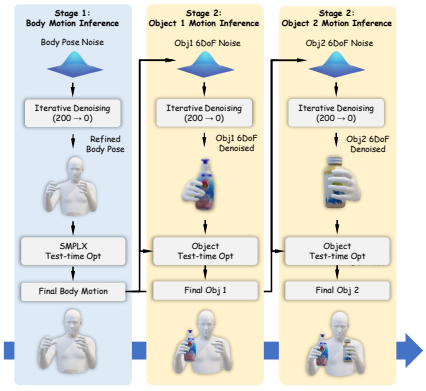

Fig. 7: Overall algorithm of diffusion inference loop: (1) Body denoising (2) 6DoF denoising.

Algorithm 2 Multi-Object Inference Loop

Require: $\mathcal{W}$, $\mathcal{H}$, $\mathbf{c}^{1:T}$, $\mathbf{m}_j^{1:T}$, $\mathbf{o}_{filtered,j}^{1:T}$, $\mathbf{M}_j$, initial states $\widehat{\mathbf{z}}_{t_d}^{1:T}$, $\{\widehat{\mathbf{o}}_{j,t_d}^{1:T}\}_{j=1}^{J}$
Ensure: Refined SMPLX Poses $\widehat{\mathbf{z}}_0^{1:T}$ and 6DoF $\{\widehat{\mathbf{o}}_{j,0}^{1:T}\}_{j=1}^{J}$

for $t_d = 200, 199, \ldots, 0$ **do**
  $\widehat{\mathbf{z}}_{t_d-1}^{1:T} \leftarrow \mathcal{W}(\widehat{\mathbf{z}}_{t_d}^{1:T}, \mathbf{c}^{1:T}, t_d)$
end for
$\widehat{\mathbf{z}}_{t_0}^{1:T} \leftarrow \mathsf{SMPLXTestTimeOpt}(\widehat{\mathbf{z}}_{t_0}^{1:T})$
for $j = 1$ **to** J **do**
  for $t_d = 200, 199, \ldots, 0$ **do**
    $\widehat{\mathbf{o}}_{j,t_d-1}^{1:T} \leftarrow \mathcal{H}(\widehat{\mathbf{o}}_{j,t_d}^{1:T}, \widehat{\mathbf{m}}_j^{1:T}, \mathbf{o}_{filtered,j}^{1:T}, t_d)$
  end for
  $\widehat{\mathbf{o}}_{j,0}^{1:T} \leftarrow \mathsf{ObjectTestTimeOpt}(\widehat{\mathbf{o}}_{j,0}^{1:T}, \widehat{\mathbf{z}}_0^{1:T}, \mathbf{M}_j)$
end for
return $\widehat{\mathbf{z}}_0^{1:T}$, $\{\widehat{\mathbf{o}}_{j,0}^{1:T}\}_{j=1}^{J}$

still allowing object inference to benefit from strong human motion priors.

More specifically, the first stage operates on the latent body pose variables $\widehat{\mathbf{z}}_{t_d}^{1:T}$, where t_d denotes the diffusion timestep and $1:T$ denotes the full temporal sequence. Starting from an initial noisy state at the largest diffusion timestep, the algorithm iteratively applies the body diffusion model until reaching the final denoised body estimate.

$$\widehat{\mathbf{z}}_{t_d-1}^{1:T} \leftarrow \mathcal{W}\big(\widehat{\mathbf{z}}_{t_d}^{1:T}, \mathbf{c}^{1:T}, t_d\big)\,, \tag{14}$$

Here, $\mathbf{c}^{1:T}$ denotes the conditioning features extracted from the central pupil frame (CPF) motion, which provide egocentric temporal cues for body reconstruction. Intuitively, this stage produces a globally consistent upper-body motion that explains the observed egomotion while remaining within the learned human motion manifold. After the diffusion sampling loop, a test-time optimization step further refines the body parameters, as described in Appendix D.

Once the body motion has been estimated, the second stage infers object motion one object at a time. For each object index $j \in \{1, \ldots, J\}$, the algorithm starts from an initial noisy object trajectory $\widehat{\mathbf{o}}_{j,t_d}^{1:T}$ and performs iterative denoising conditioned on both object-specific observations and the body motion:

$$\widehat{\mathbf{o}}_{j,t_d-1}^{1:T} \leftarrow \mathcal{H}\big(\widehat{\mathbf{o}}_{j,t_d}^{1:T}, \mathbf{m}_j^{1:T}, \mathbf{o}_{\text{filtered},j}^{1:T}, t_d\big)\,. \tag{15}$$

Here, $\mathbf{o}_{\text{filtered},j}^{1:T}$ denotes the reliable object 6DoF estimates obtained from preprocessing, and $\mathbf{m}_j^{1:T}$ denotes the object condition features, which encode the filtered object observations, coarse object translation, grasp state, and the hand-motion cues derived from the body prediction. Thus, $\mathcal{H}$ does not infer object motion from object detections alone, it completes and regularizes the object trajectory using both partial object observations and the predicted body motion.

A key property of Algorithm 2 is that multi-object interaction is handled by placing the object loop outside the HOI diffusion model. The body sequence is

inferred once and then shared across all objects, while each object trajectory is generated independently under its own observations and mesh geometry $\mathbf{M}_j$. This design preserves a common and temporally consistent human motion explanation for the entire interaction sequence. And it naturally scales to a variable number of objects without retraining a joint multi-object generator. The final outputs are the refined body motion $\widehat{\mathbf{z}}_0^{1:T}$ and refined object 6DoFs $\{\widehat{\mathbf{o}}_{j,0}^{1:T}\}_{j=1}^{J}$.

From a probabilistic perspective, the algorithm can be interpreted as a sequential conditional generation process:

$$\begin{aligned} &p\big(\widehat{\mathbf{z}}_0^{1:T}, \{\widehat{\mathbf{o}}_{j,0}^{1:T}\}_{j=1}^{J} \,\big|\, \mathbf{c}^{1:T}, \{\mathbf{m}_j^{1:T}\}_{j=1}^{J}, \{\mathbf{o}_{\text{filtered},j}^{1:T}\}_{j=1}^{J}\big) \\ &\approx p\big(\widehat{\mathbf{z}}_0^{1:T} \mid \mathbf{c}^{1:T}\big) \prod_{j=1}^{J} p\big(\widehat{\mathbf{o}}_{j,0}^{1:T} \,\big|\, \widehat{\mathbf{z}}_0^{1:T}, \mathbf{m}_j^{1:T}, \mathbf{o}_{\text{filtered},j}^{1:T}\big)\,. \end{aligned} \tag{16}$$

This factorization highlights the role of the recovered body motion as an intermediate representation: the HOI diffusion uses body motion as a strong prior to resolve object motion ambiguities. As a result, the method achieves temporally complete and physically plausible reconstruction even when object observations are sparse, noisy, or partially missing.

D Test-time Optimization

D.1 SMPL-X Test-time Optimization

Given the body diffusion prediction, we refine upper-body SMPL-X parameters by solving

$$\Theta_{\text{human}}^{\star} = \arg\min_{\Theta_{\text{human}}} \mathcal{L}_{\text{human}}, \tag{17}$$

where

$$\Theta_{\text{human}} = \big\{\theta_t^L,\, \theta_t^R,\, \theta_t^{\text{sub}}\big\}_{t=1}^{T} \cup \{\beta\}. \tag{18}$$

Here, θ_t^L and θ_t^R denote the left- and right-hand poses, θ_t^{sub} denotes the optimized upper-body subset of the SMPL-X pose, and β is a sequence-shared shape parameter. The optimized upper-body parameters are inserted back into the full SMPL-X pose, while the remaining body parameters are zeros. Global orientation and translation are computed by CPF and body poses. The objective is

$$\mathcal{L}_{\text{human}} = \lambda_{\text{body}}\mathcal{L}_{\text{body}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{kp2d}}\mathcal{L}_{\text{kp2d}} + \lambda_{\text{wrist}}\mathcal{L}_{\text{wrist}}. \tag{19}$$

For brevity, avg denotes averaging over the frames, hands, or joints for which the corresponding supervision is available.

(1) Pose anchor loss $\mathcal{L}_{\text{body}}$. Let $\hat{\theta}_t^L$, $\hat{\theta}_t^R$, and $\hat{\theta}_t^{\text{sub}}$ be the body diffusion predictions before refinement. We use

$$\mathcal{L}_{\text{body}} = \operatorname*{avg}_t\Big(\|\theta_t^{\text{sub}} - \hat{\theta}_t^{\text{sub}}\|_2^2 + \|\theta_t^L - \hat{\theta}_t^L\|_2^2 + \|\theta_t^R - \hat{\theta}_t^R\|_2^2\Big). \tag{20}$$

This term keeps the refined body configuration close to the body diffusion prediction and prevents excessive drift.

(2) Hand pose loss $\mathcal{L}_{\text{pose}}$. When WiLoR hand-pose estimates $\tilde{\theta}_t^L$ and $\tilde{\theta}_t^R$ are available, we align the refined SMPL-X hand poses to them:

$$\mathcal{L}_{\text{pose}} = \underset{t}{\text{avg}} \|\theta_t^L - \tilde{\theta}_t^L\|_2^2 + \underset{t}{\text{avg}} \|\theta_t^R - \tilde{\theta}_t^R\|_2^2. \tag{21}$$

This term regularizes the hand articulation toward the WiLoR reconstruction while preserving upper-body kinematic consistency.

(3) Hand 2D keypoint loss $\mathcal{L}_{\text{kp2d}}$. Let $\mathbf{J}_{t,j}^{\text{SMPLX}}$ and $\mathbf{J}_{t,j}^{\text{WiLoR}}$ denote the SMPL-X and WiLoR hand joints in CPF for joint j at frame t, and let $\pi_{K_t}(\cdot)$ denote perspective projection under camera intrinsics K_t. Denoting by $\mathcal{V}$ the set of jointly visible hand joints, we define

$$\mathcal{L}_{\text{kp2d}} = \underset{(t,j)\in\mathcal{V}}{\text{avg}} \left\| \pi_{K_t}\left(\mathbf{J}_{t,j}^{\text{SMPLX}}\right) - \pi_{K_t}\left(\mathbf{J}_{t,j}^{\text{WiLoR}}\right) \right\|_2^2. \tag{22}$$

This term enforces image-space consistency without requiring the optimized 3D body to match the WiLoR geometry exactly.

(4) Wrist loss $\mathcal{L}_{\text{wrist}}$. To enforce wrist 6DoF consistency in CPF, we combine positional and rotational constraints:

$$\mathcal{L}_{\text{wrist}} = \mathcal{L}_{\text{wrist}}^{\text{pos}} + \mathcal{L}_{\text{wrist}}^{\text{rot}}. \tag{23}$$

Let $\mathbf{w}_t^h$ and $\tilde{\mathbf{w}}_t^h$ denote the SMPL-X and WiLoR wrist locations in CPF for hand $h \in \{L, R\}$. The positional term is

$$\mathcal{L}_{\text{wrist}}^{\text{pos}} = \underset{t,h}{\text{avg}} \left\| \mathbf{w}_t^h - \tilde{\mathbf{w}}_t^h \right\|_2^2. \tag{24}$$

For orientation, let $R_{t,\,w\to cpf}^h \in SO(3)$ be the wrist-to-CPF rotation induced by the current SMPL-X kinematic chain, and let $\tilde{R}_{t,\,w\to cpf}^h$ be the corresponding rotation derived from WiLoR. The rotational term is

$$d_R(R_1, R_2) = \left\|\log\left(R_1 R_2^\top\right)\right\|_2, \quad \mathcal{L}_{\text{wrist}}^{\text{rot}} = \underset{t,h}{\text{avg}}\, d_R\left(R_{t,\,w\to cpf}^h, \tilde{R}_{t,\,w\to cpf}^h\right). \tag{25}$$

Together, these terms encourage physically plausible wrist motion while remaining consistent with the hand reconstruction.

D.2 Object Test-time Optimization

At test time, we perform a lightweight, fully differentiable refinement of the per-frame object 6DoF pose in CPF. For frame t, let

$$T_t^{cpf\to obj} = \left(R_t^{cpf\to obj},\, \mathbf{t}_t^{cpf\to obj}\right) \tag{26}$$

denote the object pose, and let its inverse be

$$T_t^{obj\to cpf} = \left(R_t^{obj\to cpf},\, \mathbf{u}_t^{obj\to cpf}\right) = \left(T_t^{cpf\to obj}\right)^{-1}. \tag{27}$$

Each object is optimized independently by solving

$$\Theta^{\star}_{\text{obj}} = \arg\min_{\Theta_{\text{obj}}} \mathcal{L}_{\text{object}}, \qquad \Theta_{\text{obj}} = \left\{ T_t^{cpf\to obj} \right\}_{t=1}^{T}. \tag{28}$$

The objective is

$$\begin{aligned} \mathcal{L}_{\text{object}} = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} \\ + \lambda_{\text{penetration}}\mathcal{L}_{\text{penetration}} + \lambda_{\text{temporal}}\mathcal{L}_{\text{temporal}}. \end{aligned} \tag{29}$$

with

$$\mathcal{L}_{\text{contact}} = \mathcal{L}^{\text{prox}}_{\text{contact}} + \mathcal{L}^{\text{noslip}}_{\text{contact}}. \tag{30}$$

Thus, consistent with the main paper, the contact term combines a surface-proximity term and a no-slip term. Again, avg denotes averaging over the indices for which the corresponding quantities are defined.

To measure object-surface proximity, let $\mathcal{V}_{\mathcal{O}}$ be the object mesh vertices in the object local frame and define the nearest-neighbor distance proxy

$$\tilde{d}_{\mathcal{O}}(\mathbf{x}) = \min_{\mathbf{v}\in\mathcal{V}_{\mathcal{O}}} \|\mathbf{x}-\mathbf{v}\|_2. \tag{31}$$

(1) Object anchor loss $\mathcal{L}_{\text{obj}}$. Let

$$T_t^{obj\to cpf,(0)} = \left(R_t^{obj\to cpf,(0)},\, \mathbf{u}_t^{obj\to cpf,(0)} \right) \tag{32}$$

be the HOI Diffusion prediction before refinement. We regularize the refined object pose toward this initialization:

$$\mathcal{L}_{\text{obj}} = \operatorname*{avg}_t \Big(d_R\Big(R_t^{obj\to cpf}, R_t^{obj\to cpf,(0)} \Big)^2 + \Big\| \mathbf{u}_t^{obj\to cpf} - \mathbf{u}_t^{obj\to cpf,(0)} \Big\|_2^2 \Big). \tag{33}$$

This term prevents excessive drift from the HOI Diffusion prediction.

(2) Contact loss $\mathcal{L}_{\text{contact}}$.

Surface proximity. Let $\mathcal{P}_t$ be the set of palm sample points from the grasping hand at frame t. The proximity term encourages these samples to stay close to a target distance d_0 from the object surface:

$$\mathcal{L}^{\text{prox}}_{\text{contact}} = \operatorname*{avg}_{t,\,\mathbf{p}\in\mathcal{P}_t} \Big[\tilde{d}_{\mathcal{O}}\Big(R_t^{cpf\to obj}\mathbf{p} + \mathbf{t}_t^{cpf\to obj} \Big) - d_0 \Big]_+^2, \tag{34}$$

where $[x]_+ = \max(x, 0)$.

No-slip contact consistency. Let $\mathbf{h}_{t,i}^{cpf}$ be the i-th vertex of the grasping hand in CPF, and let

$$\mathbf{h}_{t,i}^{obj} = R_t^{cpf\to obj}\mathbf{h}_{t,i}^{cpf} + \mathbf{t}_t^{cpf\to obj} \tag{35}$$

be the same vertex expressed in the object local frame. For a temporal window $k = 1, \ldots, K_c$, we consider neighboring frames with the same grasping hand and penalize the motion of these hand vertices in the object local frame:

$$\mathcal{L}^{\text{noslip}}_{\text{contact}} = \sum_{k=1}^{K_c} \operatorname*{avg}_{(t,t+k),\,i} w_{t,i}^{(k)} \left\| \frac{\mathbf{h}_{t+k,i}^{obj} - \mathbf{h}_{t,i}^{obj}}{k} \right\|_2^2, \tag{36}$$

where $w_{t,i}^{(k)} \in [0,1]$ is a soft contact weight that increases as the vertex gets closer to the object surface. This term directly encodes the near-zero relative motion assumption at contact and suppresses unnatural sliding.

(3) Penetration loss $\mathcal{L}_{\text{penetration}}$. Let $\mathcal{B}$ denote sampled points from both hands. We penalize points that violate a safety margin δ to the object surface:

$$\mathcal{L}_{\text{penetration}} = \operatorname*{avg}_{t,\,\mathbf{q}\in\mathcal{B}} \Big[\delta - \tilde{d}_{\mathcal{O}}\Big(R_t^{cpf\to obj}\mathbf{q} + \mathbf{t}_t^{cpf\to obj}\Big)\Big]_+^2. \tag{37}$$

This term discourages non-physical interpenetration.

(4) Temporal loss $\mathcal{L}_{\text{temporal}}$. To encourage smooth object motion, we regularize both velocity and acceleration of the inverse pose $T_t^{obj\to cpf}$. For $k = 1, \ldots, K_s$, we define

$$\begin{aligned} \boldsymbol{\omega}_t^{(k)} &= \frac{1}{k}\log\!\Big(R_{t+k}^{obj\to cpf}\left(R_t^{obj\to cpf}\right)^{\top}\Big), & \mathbf{v}_t^{(k)} &= \frac{\mathbf{u}_{t+k}^{obj\to cpf} - \mathbf{u}_t^{obj\to cpf}}{k}, \\ \dot{\boldsymbol{\omega}}_t^{(k)} &= \frac{\boldsymbol{\omega}_{t+k}^{(k)} - \boldsymbol{\omega}_t^{(k)}}{k}, & \mathbf{a}_t^{(k)} &= \frac{\mathbf{v}_{t+k}^{(k)} - \mathbf{v}_t^{(k)}}{k}. \end{aligned} \tag{38}$$

The temporal loss is then

$$\begin{aligned} \mathcal{L}_{\text{temporal}} = \sum_{k=1}^{K_s}\Big(&\alpha_v \operatorname*{avg}_t\big(\|\boldsymbol{\omega}_t^{(k)}\|_2^2 + \|\mathbf{v}_t^{(k)}\|_2^2\big) \\ &+ \alpha_a \operatorname*{avg}_t\big(\|\dot{\boldsymbol{\omega}}_t^{(k)}\|_2^2 + \|\mathbf{a}_t^{(k)}\|_2^2\big)\Big). \end{aligned} \tag{39}$$

This term suppresses unstable rotational and translational changes and improves trajectory smoothness.