EgoHOI: Prior-Guided 3D Hand-Object Interaction Reconstruction from Monocular Egocentric RGB Video

Zixuan Wang
Department of Physics
Tsinghua University
Beijing 100084, China
wang-zx23@mails.tsinghua.edu.cn

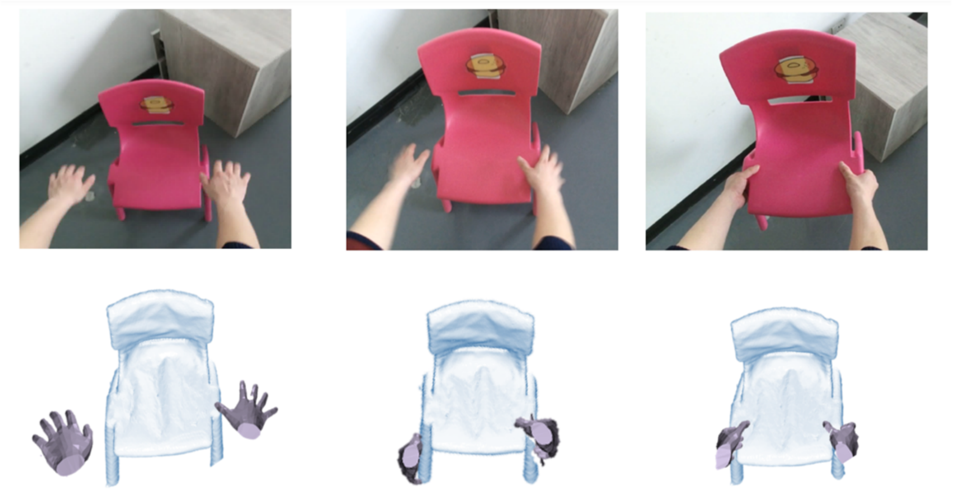


Figure 1: **Overview of EgoHOI.** Our framework takes a raw, egocentric RGB video capturing a hand-object interaction process (upside) and produces a physically plausible, temporally consistent 3D reconstruction (downside). We successfully resolve occlusions and refine noisy initial estimates into smooth, penetration-free motion trajectories by integrating differentiable physical priors into an optimization process.

1 Introduction

The reconstruction of 3D hand-object interactions (HOI) from monocular egocentric videos is a pivotal task with far-reaching implications for robotics, augmented reality, and human behavior analysis. Despite its potential, accurately capturing the dynamic interplay between hands and objects from a single RGB viewpoint remains a formidable challenge. The core difficulties stem from a confluence of factors: persistent object occlusion by the hand, the inherent noise and inaccuracies of upstream perception models, and the frequent violation of fundamental physical laws in the resulting reconstructions. Consequently, existing methodologies often yield 3D trajectories marred by artifacts such as inter-penetration, unnatural hand articulations, and temporal discontinuities. These deficiencies largely arise from an over-reliance on imperfect initial estimates from hand pose regressors or object detectors, a failure to explicitly model the dynamics of the interaction sequence, and a lack of robust mechanisms to enforce physical plausibility.

To overcome these limitations, we introduce **EgoHOI**, a novel neural optimization framework designed to refine noisy initial estimates into physically coherent and realistic 3D HOI trajectories. Our approach is engineered to systematically address the aforementioned challenges by uniquely integrating three core principles. First, we employ an occlusion-aware processing pipeline that leverages large language model (LLM) powered segmentation and advanced video inpainting to robustly estimate object properties even under severe occlusion. Second, we formulate a set of differentiable physical constraints, encoding principles such as contact maintenance and non-penetration directly into the optimization objectives. Third, our framework operates in a self-supervised manner, utilizing 2D-3D consistency to refine trajectories without requiring any 3D ground truth annotations. This design ensures our method is not only robust to occlusion and physically grounded but also highly practical, scaling effectively for use with consumer-grade capture devices like smartphone cameras.

2 Methodology

2.1 Problem Formulation

Given a monocular egocentric RGB video sequence $\mathcal{V} = \{I_t\}_{t=1}^{T}$, our method takes as input initial, noisy reconstructions from upstream models. These include MANO hand parameters $\mathbf{H}_{\text{init}} = \{\mathbf{h}_t\}_{t=1}^{T}$, object 6D poses $\mathbf{O}_{\text{init}} = \{\mathbf{o}_t\}_{t=1}^{T}$, and camera poses from SLAM $\mathbf{C}_{\text{init}} = \{\mathbf{c}_t\}_{t=1}^{T}$. The objective is to jointly optimize the hand and object trajectories, represented by a set of time-varying parameters $\mathbf{\Theta} = (\mathbf{H}, \mathbf{O})$. We seek the optimal parameter set $\mathbf{\Theta}^*$ that minimizes a comprehensive, prior-guided loss function $\mathcal{L}$, formally stated as:

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}; \mathbf{\Theta}_{\text{init}}, \mathcal{V}) \tag{1}$$

The optimized output $\mathbf{\Theta}^*$ represents a physically plausible and temporally coherent 3D reconstruction of the hand-object interaction.

2.2 Occlusion-Aware Initialization Pipeline

Our framework, depicted in Figure 2(a), begins with a two-stage process. The first stage focuses on robustly initializing the hand, object, and camera trajectories from the input video. We employ the pre-trained HaWoR model [1] to regress initial hand parameters in the MANO format and concurrently estimate the camera trajectory via SLAM. A critical challenge in egocentric HOI is estimating the 6D pose of the object while it is heavily occluded by the hand. To address this, we introduce an occlusion-aware object perception module, shown in Figure 2(b). This module first utilizes LISA++ [2] for reasoning-based segmentation to isolate the visible parts of the hand and object. Subsequently, we leverage ProPainter [3] to perform video inpainting on the occluded regions, effectively generating a "clean" video of the object without the hand. From this inpainted video, we can reliably estimate a 6D object pose of the onjects and obtain a consistent mesh representation using the SOTA method FoundationPose [4]. This initialization stage provides a strong, albeit still noisy, starting point for the subsequent refinement.

2.3 Prior-Guided Optimization Objectives

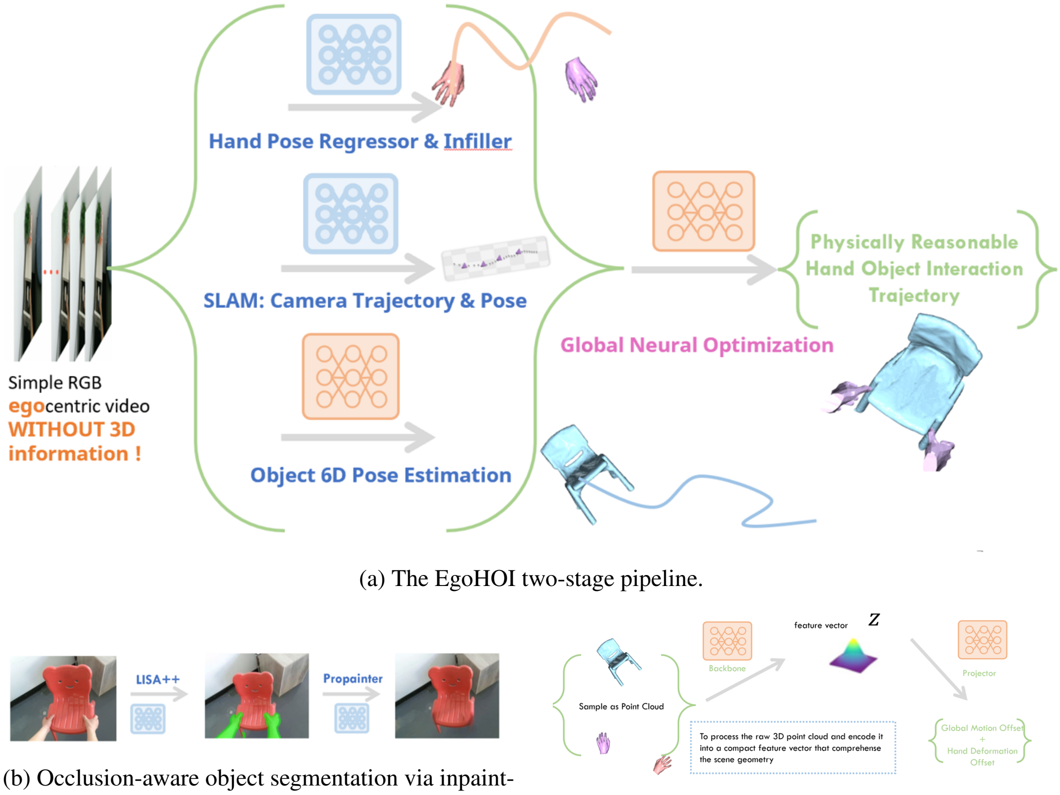
The second stage of our framework refines the initial trajectories through a neural optimization process, as illustrated in Figure 2(c). This is guided by a composite loss function $\mathcal{L}$ that encodes strong physical and geometric priors. The total loss is a weighted sum of five differentiable terms:

$$\mathcal{L} = \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{imitation}}\mathcal{L}_{\text{imitation}} + \lambda_{\text{reproj}}\mathcal{L}_{\text{reproj}} \tag{2}$$

To enforce physical contact, the **contact loss** $\mathcal{L}_{\text{contact}}$ encourages proximity between the hand and object surfaces. It is defined as the average minimum distance from points sampled on the hand surface to the object surface:

$$\mathcal{L}_{\text{contact}} = \frac{1}{T \cdot K} \sum_{t=1}^{T} \sum_{k=1}^{K} \min_{q \in S_O(t)} \|p_k(t) - q\|_2$$

where $S_H(t)$ and $S_O(t)$ are the hand and object surfaces, and $\{p_k(t)\}_{k=1}^{K}$ are points sampled from $S_H(t)$.



To process the raw 3D point cloud and encode it into a compact feature vector that comprehense the scene geometry

(a) The EgoHOI two-stage pipeline.

(b) Occlusion-aware object segmentation via inpainting.

(c) Gradient-based refinement using physical priors.

Figure 2: **Detailed Framework of EgoHOI.** (a) Our method consists of two main stages: an occlusion-aware initialization stage that leverages pre-trained models for initial pose estimation, and a neural optimization stage that refines the trajectories. (b) To handle severe hand-object occlusion, we first segment the hand and then apply video inpainting to reconstruct the object's appearance, enabling a more robust 6D pose estimation. (c) The core of our method is a gradient-based optimization process guided by a composite loss function that enforces physical plausibility, temporal smoothness, and 2D-3D consistency.

To prevent physically impossible intersections, the **penetration loss** $\mathcal{L}_{\text{pen}}$ penalizes any hand vertices that lie inside the object mesh. This is efficiently computed using the object's signed distance function (SDF) ϕ_{O_t}:

$$\mathcal{L}_{\text{pen}} = \frac{1}{T \cdot N_v} \sum_{t=1}^{T} \sum_{j=1}^{N_v} \max(-\phi_{O_t}(v_{j,t}^H), 0)^2$$

where $v_{j,t}^H$ is the j-th hand vertex at time t, and $\phi_{O_t}(v) < 0$ for points inside the object.

To ensure natural and fluid motion, the **temporal smoothness loss** $\mathcal{L}_{\text{smooth}}$ minimizes the acceleration of hand vertices across time, approximated using a second-order finite difference:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{(T-2) \cdot N_v} \sum_{t=2}^{T-1} \sum_{j=1}^{N_v} \|(v_{j,t+1}^H - v_{j,t}^H) - (v_{j,t}^H - v_{j,t-1}^H)\|_2^2$$

To preserve the overall intent of the initial motion while correcting its flaws, the **imitation loss** $\mathcal{L}_{\text{imitation}}$ regularizes the optimized parameters θ_t to not deviate excessively from their initial estimates $\theta_t^{\text{initial}}$:

$$\mathcal{L}_{\text{imitation}} = \frac{1}{T \cdot D_\theta} \sum_{t=1}^{T} \left\|\theta_t - \theta_t^{\text{initial}}\right\|_2^2$$

Finally, to anchor the 3D reconstruction to the 2D evidence, the **reprojection loss** $\mathcal{L}_{\text{reproj}}$ enforces consistency between the rendered semantic masks of the 3D models and the ground truth 2D masks derived from the input video:

$$\mathcal{L}_{\text{reproj}} = \sum_{t=1}^{T} \|M_t - \tilde{M}_t\|_1$$

where M_t is the rendered mask and $\tilde{M}_t$ is the observed mask. Collectively, these gradient-based priors guide the optimization to a physically plausible and accurate reconstruction.

3 Results and Discussion

Our experimental evaluation demonstrates the effectiveness of EgoHOI in producing high-fidelity 3D reconstructions of hand-object interactions. As shown in Figure 3, our method successfully corrects common failure modes of existing approaches. We compare our refined outputs against the noisy initial estimates, showing significant improvements in physical plausibility, contact accuracy, and temporal smoothness.

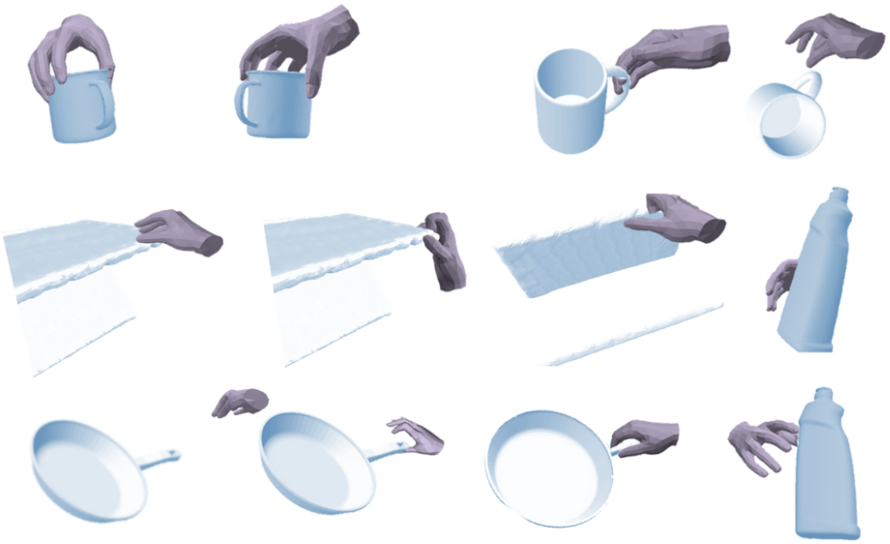


Figure 3: **Qualitative Results.** Our method resolves inter-penetration issues, improves contact realism between the hand and the object, and generates a temporally smooth motion trajectory.

References

[1] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos, 2025.

[2] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model, 2024.

[3] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting, 2023.

[4] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects, 2024.