AlignHOI: Hand–Object Reconstruction via Alignment and Refinement

Liting Wen[1] Xin Lv[2] Jungbin Cho[1,3] László A. Jeni[1] Xiao-Xiao Long[2]
[1] Carnegie Mellon University [2]Nanjing University [3]Yonsei University

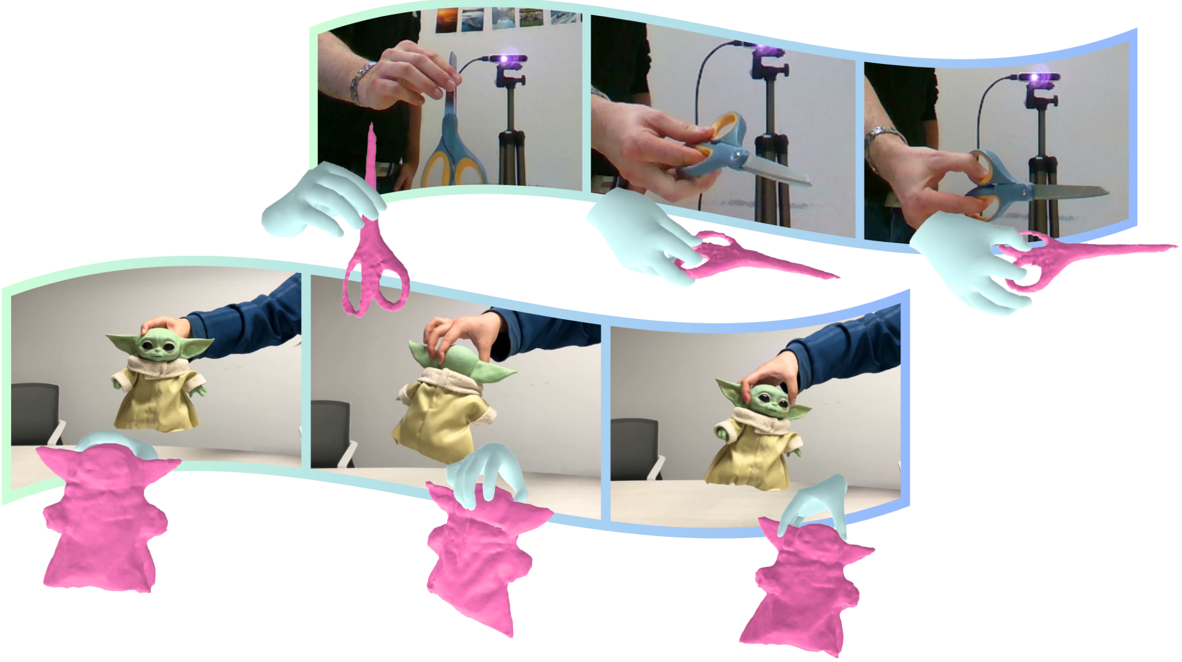

Figure 1. AlignHOI reconstructs accurate and coherent hand–object interactions from monocular video via alignment and refinement. This unified strategy enables high-fidelity object reconstruction and stable, realistic 3D interactions across diverse manipulation sequences.

Abstract

Reconstructing Hand–Object Interactions (HOI) is essential for understanding human manipulation and enabling downstream applications such as AR/VR and robotics. However, accurate HOI reconstruction from videos remains challenging: in-hand object pose estimation is ill-posed due to unknown object shape and severe occlusions, while the reconstruction of hand–object interaction from disparate coordinate systems suffers from scale and depth ambiguities that hinder accurate interaction modeling. We observe that these challenges share a common root: the underlying solutions remain fundamentally ambiguous when inferred from ill-posed monocular observations. Motivated by this insight, we introduce ***AlignHOI****, a unified alignment–refinement framework that converts these ambiguous subproblems into well-constrained optimization tasks. Within AlignHOI, both object pose estimation and hand reconstruction follow the same principle: coarse alignment first restricts the solution space, and the remaining discrepancies are then resolved through refinement. Specifically, for object pose estimation, we cast the problem as finite template matching by enumerating rendered 3D pose templates derived from a generated prior and retrieving the plausible candidate, thereby converting continuous estimation into a tractable matching problem. The retrieved pose is then further refined to bridge the gap between the discrete template space and the continuous pose domain. For hand–object interaction, we align the hand to the object using their 3D relative position to address scale and depth ambiguities, followed by refinement to enforce physically plausible interaction. Experiments demonstrate that AlignHOI achieves state-of-the-art accuracy and reconstruction quality while running significantly faster than existing methods.*

1. Introduction

Reconstructing hand–object interactions involves recovering the 3D geometry and pose of both hands and objects from visual observations such as videos or images. This task is crucial for understanding how humans manipulate objects and interact with the physical world, and it further supports a range of popular applications such as in-hand object scanning [17], robotic manipulation [33], and AR/VR [1].

A key challenge in hand–object reconstruction is how to obtain accurate object poses for previously unseen objects. This problem is particularly significant and difficult to address due to two main reasons: (1) **Lack of a universal prior**. Unlike hands, where strong statistical models provide useful priors, objects are highly diverse and category-agnostic, making it difficult to define a unified representation for pose estimation. (2) **Frequent occlusions** during interactions. When a hand grasps or manipulates an object, the object is often partially covered by the hand, making it challenging to acquire sufficient visual evidence for reliable pose estimation. In particular, such occlusions lead to too few reliable feature correspondences, which is a common failure case for reconstruction methods that rely on structure-from-motion (SfM) initialization.

Several representative works reflect the challenge discussed above. Huang et al. [16] leverages hand priors [29] to obtain accurate hand poses and directly uses them as object poses. However, the method assumes no relative shift between the hand and the object. This strong assumption prohibits dynamic hand–object interactions, as it requires the hand to grasp the object at the same location throughout the entire sequence. HOLD [10] and MagicHOI [40] rely on SfM for initialization. Since SfM struggles under occlusion and with textureless objects, these methods usually incorporate additional strategies to mitigate the effect of inaccurate poses. However, such strategies only partially alleviate the problem and do not fundamentally resolve the issue of unreliable pose estimation in hand-object reconstruction.

In this paper, we propose **AlignHOI**, a novel framework for reconstructing hand–object interactions using an alignment–refinement strategy. AlignHOI achieves both accurate object pose estimation and high-quality mesh recovery under severe occlusions and dynamic interactive motion. We reformulate object pose estimation as a finite template-matching problem, where rendered pose templates are first retrieved to coarsely align the object pose and then refined via optimization, enabling stable and efficient estimation even under partial occlusions. Once the pose is estimated, we perform implicit surface reconstruction to obtain fast and high-fidelity object geometry and appearance.

Even with accurate object shape and pose, hand–object misalignment commonly occurs due to inherent scale and depth ambiguities. To address this, we first align the hand to the object by predicting their 3D relative position, and then refine the hand reconstruction to ensure physically plausible interaction. Together, these components enable robust hand–object reconstruction with state-of-the-art efficiency and accuracy.

Our key insight is to use geometric priors for coarse pose alignment followed by optimization-based refinement. For object pose estimation, we cast the open-ended 3D prediction problem as finite template matching, enabling stable and efficient estimation under occlusions. For hand–object interaction, we align the hand to the object via their 3D relative position to resolve scale and depth ambiguities, and refine the result using physical and geometric constraints for realistic interaction.

In summary, our contributions are:

- We propose a stable and effective formulation for category-agnostic object pose estimation under severe in-hand occlusions. By converting the ill-posed continuous pose regression problem into a tractable template-matching problem, our method enables reliable pose recovery in challenging hand–object interaction scenarios.
- We develop a highly efficient hand–object reconstruction pipeline that delivers fast and high-quality 3D reconstruction across long interaction sequences. Specifically, we extend an efficient neural representation framework to support dynamic HOI reconstruction, enabling significantly faster processing compared to existing methods.
- We introduce a HOI optimization strategy that is both spatially consistent and physically plausible, effectively resolving cross-coordinate inconsistencies between the hand and the object. Extensive experiments demonstrate that our method achieves stable and coherent 3D alignment across diverse interaction scenarios.

2. Related Works

Object reconstruction: Recent advances in 3D reconstruction from 2D images have made remarkable progress. Classical methods rely on multi-view stereo [11, 19], while recent works employ neural representations [37] for more accurate reconstructions. However, most still depend on SfM [4], which are not accurate under dynamic, low-textured scenes [16]. To address this, data-driven approaches leverage large-scale datasets [9] and train feed-forward networks [26, 35] to infer 3D directly from images. Our approach leverages advantages from both neural representations [41] and feed-forward models [43] to tackle a much more challenging task of reconstructing not only objects but also hands from monocular videos.

Hand-held object reconstruction: Reconstructing hand-held objects has gained increasing attention in recent years, yet remains highly challenging due to frequent hand occlusions and the wide variety of unseen object categories.

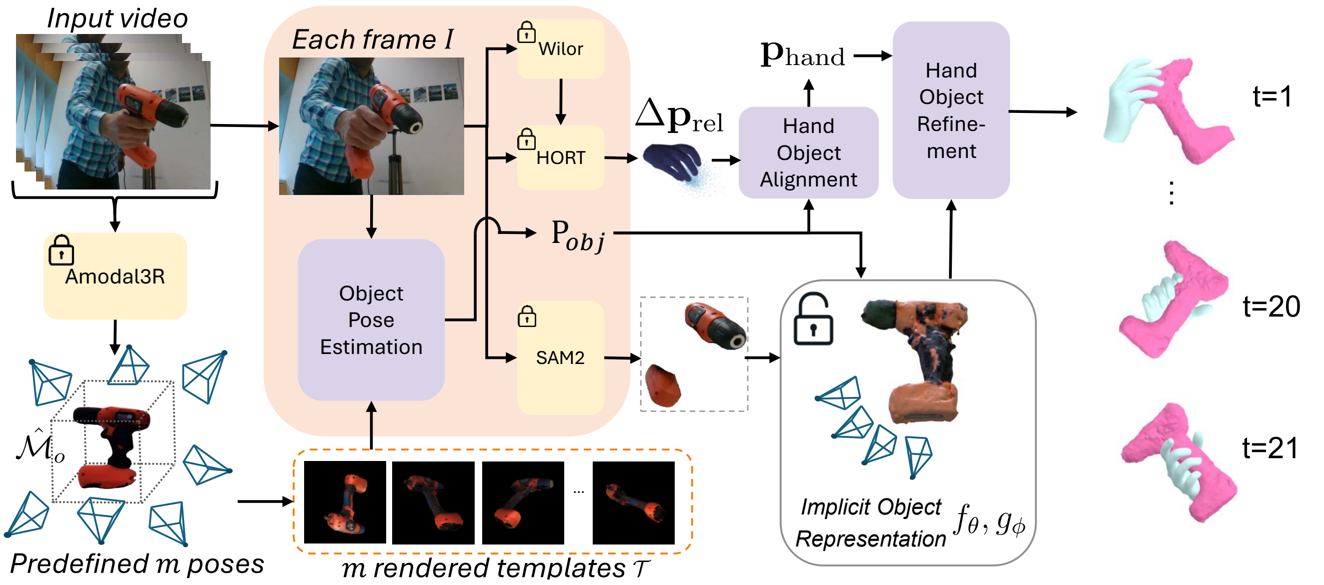

Figure 2. **Pipeline overview.** We first estimate in-hand object poses via template matching, followed by geometric-consistency verification and continuous refinement. Using these poses, we reconstruct the object with an implicit SDF representation parameterized by f_θ and g_ϕ. Finally, we align the reconstructed object and the hand within a unified coordinate system using HORT, and refine the result through spatially and physically consistent hand–object optimization. This enforces accurate alignment and realistic interaction, enabling robust reconstruction under occlusions, low-texture objects, and dynamic hand–object motion.

Many approaches assume that the object shape is known and focus on estimating its pose [4, 44]. While effective, ground-truth object templates are typically unavailable in real-world settings. Learning-based methods [15, 18] attempt to predict both object shape and poses directly, but they generalize poorly to unseen object shapes. Consequently, recent works explore neural representations [37] to improve object shape generalization. Huang et al. [16] leverage the MANO hand prior [29] to approximate object poses in static settings. BundleSDF [42] and Hampali et al. [14] extend reconstruction to dynamic interactions by jointly optimizing the object mesh and its pose trajectory. HOLD [10] and MagicHOI [40] further improve dynamic reconstruction quality, but rely on SfM initialization, which is unstable under severe occlusions. Chen et al. [5] track an implicit object shape by aligning per-frame point clouds to its SDF surface, but require access to the object point cloud at every frame. Most related to our work, Jiang et al. [17] leverage a text-to-3D generative model to synthesize a pseudo object template, reducing the difficulty of reconstructing unknown objects. However, optimizing the pseudo-template pose in continuous 6-DoF space is inherently unstable and often results in degraded reconstructions. We instead formulate this step as a discrete template-matching problem, followed by a refinement stage for accurate pose estimation in continuous space.

Hand-object interaction reconstruction: Beyond accurate object shapes and poses, hand-object interaction reconstruction additionally aims to capture the geometric relationship between object and hands in the 3D space [4, 6, 8, 36, 44, 46]. While prior works suffer from inconsistent initialization of hand and object coordinate systems, methods such as HOLD [10] and EasyHOI [21] tackle this issue using 2D mask supervision for global alignment followed by refinement. However, mask-based alignment enforces only silhouette overlap, often producing hand–object pairs that appear aligned in the 2D camera view but remain misaligned in 3D space. We observe that the relative hand–object relationship remains invariant across coordinate systems and can therefore serve as a geometric prior. We leverage a feed-forward method [7] to estimate this relationship, aligning the hand parameters within the object coordinate frame for consistent interaction reconstruction.

3. Method

Given a monocular input video, our objective is to accurately reconstruct both the shape and the pose of an interacting hand and object. Our method consists of three stages. (1) **Object pose estimation** (§ 3.2), where a finite template-matching formulation enumerates pose hypotheses through rendered templates. Combining feature-based retrieval, geometric-consistency selection, and continuous pose refinement, this stage remains stable and accurate even for thin objects, textureless surfaces, and severe hand–object occlusions (see Fig. 4). (2) **Fast object reconstruction** (§ 3.3), where we efficiently recover object mesh using the estimated poses. (3) **Hand–object interaction optimization** (§ 3.4), where we refine hand pose and hand–object alignment to obtain accurate interactions.

3.1. Preparation

Coarse Object Prior for Template Matching: We use Amodal3R [43] to generate an initial object mesh $\hat{\mathcal{M}}_o$ for rendering template views in § 3.2. Note that we only use this mesh as a geometric prior in the object pose matching stage, where it provides 3D-consistent renderings for template retrieval. Because our in-hand object pose estimation strategy is robust to geometric variations and texture differences, this generated prior is fully sufficient for reliable initialization and ensures that subsequent optimization starts from a geometry-faithful and occlusion-robust pose.

Hand initialization: For each frame, we use an off-the-shelf hand model [27] to estimate MANO parameters [29], including hand pose $\theta \in \mathbb{R}^{45}$, hand shape $\beta \in \mathbb{R}^{10}$, global rotation $R_h \in SO(3)$, and translation $\mathbf{t}_h \in \mathbb{R}^3$.

3.2. Object Pose Estimation

Given $\hat{\mathcal{M}}_o$, we render it from m different predefined object poses $\mathcal{P} = \{\mathbf{p}_i^{\text{temp}}\}_{i=1}^{m}$ to build a template set of images $\mathcal{T} = \{T_i\}_{i=1}^{m}$. For each video frame I, we identify the most similar template image T_{win} and use its associated canonical pose $\mathbf{p}_{\text{win}}^{\text{temp}}$ as a coarse initialization of the object pose. We then refine this initialization in continuous pose space to close the quantization gap introduced by the discrete template poses. By constraining the search to a discrete set of poses, we convert an unbounded prediction problem into a finite, well-behaved search space, enabling stable and efficient object pose estimation.

Object Pose Matching: Our goal is to retrieve the object template whose predefined pose best matches the observed video frame. We begin by extracting patch-level features from each template $T_i \in \mathcal{T}$ using DINOv2. Following [25], all template features are clustered via k-means, and each template is encoded as a k-dimensional cluster-based vector representation.

For each video frame I, we compute the cosine similarity between its feature embedding and each template descriptor,

$$s(I, T_i) = \frac{\langle f(I), f(T_i) \rangle}{\|f(I)\| \, \|f(T_i)\|}, \tag{1}$$

where $f(\cdot)$ denotes the feature embedding. Since cosine similarity normalizes vectors and emphasizes their directions rather than magnitudes, the matching becomes insensitive to the number of visible features and thus remains robust even under partial occlusion.

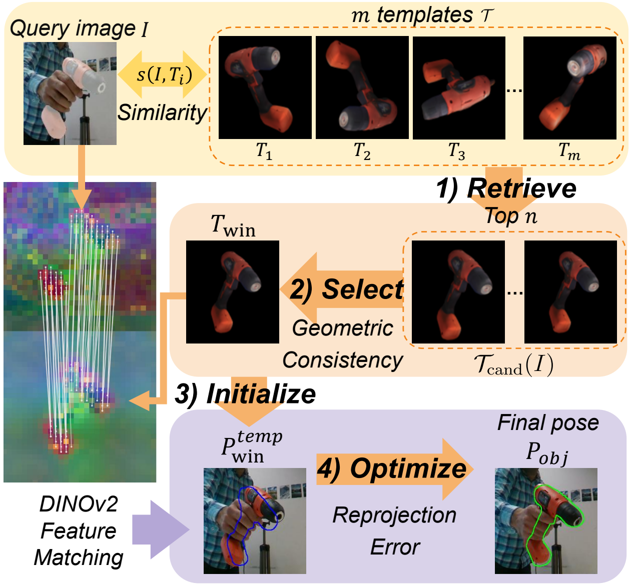

Figure 3. **Object Pose Estimation.** Given a query image, we **1) Retrieve** top-n candidate templates that have predefined poses, and then **2) Select** the winner template using geometric consistency. Using the winner template pose as an **3) Initialization**, we **4) Optimize** the pose to obtain final pose estimation.

While this coarse matching is efficient, we observe that the top-ranked template does not always provide the most accurate geometric alignment, as feature similarity alone may overlook fine-grained geometric cues. To mitigate this, we first retain the top-n templates with the highest similarity scores:

$$\mathcal{T}_{\text{cand}}(I) = \text{Top-}n\big(\{\, s(I, T_i) \,\}_{i=1}^{m}\big), \tag{2}$$

where $\text{Top-}n(\cdot)$ returns the n templates with the highest similarity scores. We then assess geometric consistency within this candidate set by estimating a pose for each template via RANSAC-PnP and selecting the template whose estimated pose yields the largest inlier count. This inlier-based selection complements the feature-based retrieval, leading to accurate and stable template matching even under heavy occlusions and complex hand–object interactions.

Object Pose Refinement: The retrieved template pose $\mathbf{p}_{\text{win}}^{\text{temp}}$, which is selected from a discrete pose set $\mathcal{P}$, provides only a coarse approximation of the true object pose. To obtain a continuous and accurate estimate, we match DINOv2 features between the query image and the retrieved template T_{win}. Using these correspondences, we then refine the pose $\mathbf{p}_{\text{win}}^{\text{temp}}$ via iterative PnP optimization, minimizing the reprojection error and producing a pose $\mathbf{P}_{\text{obj}}$ that is geometrically consistent with the input image. Notably, all poses are estimated independently per frame, and we do not apply additional temporal smoothing, since the method already achieves stable results.

3.3. Efficient Implicit Object Reconstruction

Once the object pose is estimated, we reconstruct the object geometry and appearance using an implicit neural representation. The geometry is modeled with a signed distance field (SDF) network, while the appearance is produced by a companion color network. Formally, the geometry branch is a function

$$f_\theta : (\mathbf{x}, \mathbf{c}) \mapsto (s, \mathbf{z}), \tag{3}$$

which encodes a 3D point $\mathbf{x} \in \mathbb{R}^3$ together with a conditioning code $\mathbf{c}$ that models appearance changes, and outputs its signed distance $s \in \mathbb{R}$ along with a latent feature vector $\mathbf{z}$. The color branch is another function

$$g_\phi : (\mathbf{x}, \mathbf{n}, \mathbf{z}) \mapsto \mathbf{c}_{rgb}, \tag{4}$$

that maps the latent features $\mathbf{z}$, 3D position $\mathbf{x}$, and estimated surface normal $\mathbf{n}$ to an RGB color $\mathbf{c}_{rgb} \in [0, 1]^3$. These two branches are optimized end-to-end via differentiable volume rendering.

Notably, we integrate the instant surface reconstruction strategy of [41] into our dynamic HOI pipeline, enabling highly efficient rigid object reconstruction under hand motion. For dense sequences with several hundred frames, our method completes reconstruction within only 16 minutes. Compared to Jiang et al. [17], our approach is **30×** faster, and compared to HOLD [10], it achieves an impressive **112×** speedup. We provide additional technical clarifications regarding our efficient implicit object reconstruction in the *Supplementary Material*.

3.4. Hand-Object Interaction Optimization

We aim to produce hand–object interactions that are both spatially consistent and physically plausible. A learned 3D relative offset ensures accurate hand placement, while physically motivated constraints refine the interaction to encourage contact and prevent penetration.

Hand-Object Alignment: Since the object and hand are estimated in different coordinate systems, explicit alignment is required. Prior works [10, 21] employ 2D mask supervision by matching rendered masks with off-the-shelf segmentations, yet such supervision is insufficient to guarantee geometrically consistent alignment in 3D space. Specifically, the hand and object may appear well-aligned in 2D projections while remaining spatially distant in 3D due to inherent depth and scale ambiguities. Our key observation is that the relative spatial relationship between the hand and the object remains consistent across coordinate systems, enabling us to use it as a geometric prior to align the hand within the object coordinate space. Building on this observation, we employ HORT [7] to estimate the 3D relative offset $\Delta\mathbf{p}_{\text{rel}}$ between the hand and object. The estimated offset serves as a geometric prior that aligns hand parameters from the off-the-shelf model [27] within the object coordinate system:

$$\mathbf{p}_{\text{hand}} = \mathbf{p}_{\text{obj}} + \left(\mathbf{p}_{\text{hand}}^{\text{HORT}} - \mathbf{p}_{\text{obj}}^{\text{HORT}}\right) = \mathbf{p}_{\text{obj}} + \Delta\mathbf{p}_{\text{rel}}. \tag{5}$$

Here, $\mathbf{p}_{\text{obj}}$ denotes the object position in its coordinate system, while $\mathbf{p}_{\text{hand}}^{\text{HORT}}$ and $\mathbf{p}_{\text{obj}}^{\text{HORT}}$ represent the hand and object positions in the HORT coordinate system, respectively, from which we obtain $\Delta\mathbf{p}_{\text{rel}}$. These terms together define the aligned hand position $\mathbf{p}_{\text{hand}}$ in the object frame. This formulation provides a simple yet effective way to ensure spatially consistent hand–object alignment.

Hand-Object Interaction Refinement: We further refine the hand parameters to ensure physically plausible interaction with the object. We optimize the hand translation $\mathbf{t}_h$ and shape parameters $\boldsymbol{\beta}$ under several physically motivated constraints. Specifically, a mask IoU loss $\mathcal{L}_{\text{mask}}$ enforces silhouette consistency between rendered masks and off-the-shelf segmentations [28], a contact loss $\mathcal{L}_{\text{contact}}$ minimizes the distance between hand contact points and the object surface following [10], and a penetration loss $\mathcal{L}_{\text{pene}}$ penalizes penetrations between hand vertices and the object mesh:

$$\mathcal{L}_{\text{ref}} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{pene}}\mathcal{L}_{\text{pene}}, \tag{6}$$

where λ_{mask}, λ_{contact}, and λ_{pene} denote the corresponding loss weights. This refinement effectively reduces hand–object penetration and improves the geometric consistency of the reconstructed interaction. See more details in the *Supplementary Material*.

4. Experiments

4.1. Implementation Details

We use SAM 2 [28] to obtain both object and hand masks, providing clean and reliable segmentation cues for pose estimation and HOI optimization. For object pose estimation, we extract DINOv2 [24] ViT-g/14 features from the 30th layer. We empirically find that this intermediate layer offers an effective balance between local geometric detail and global semantic context, which is crucial for robust template matching under severe occlusions and low-texture conditions. During training, it is worth noting that we extend the efficient computation framework [41] to the dynamic HOI reconstruction pipeline, which enables processing a sequence of 400 frames on a single RTX 4090 GPU in approximately 16 minutes. More details are provided in the *Supplementary Material*.

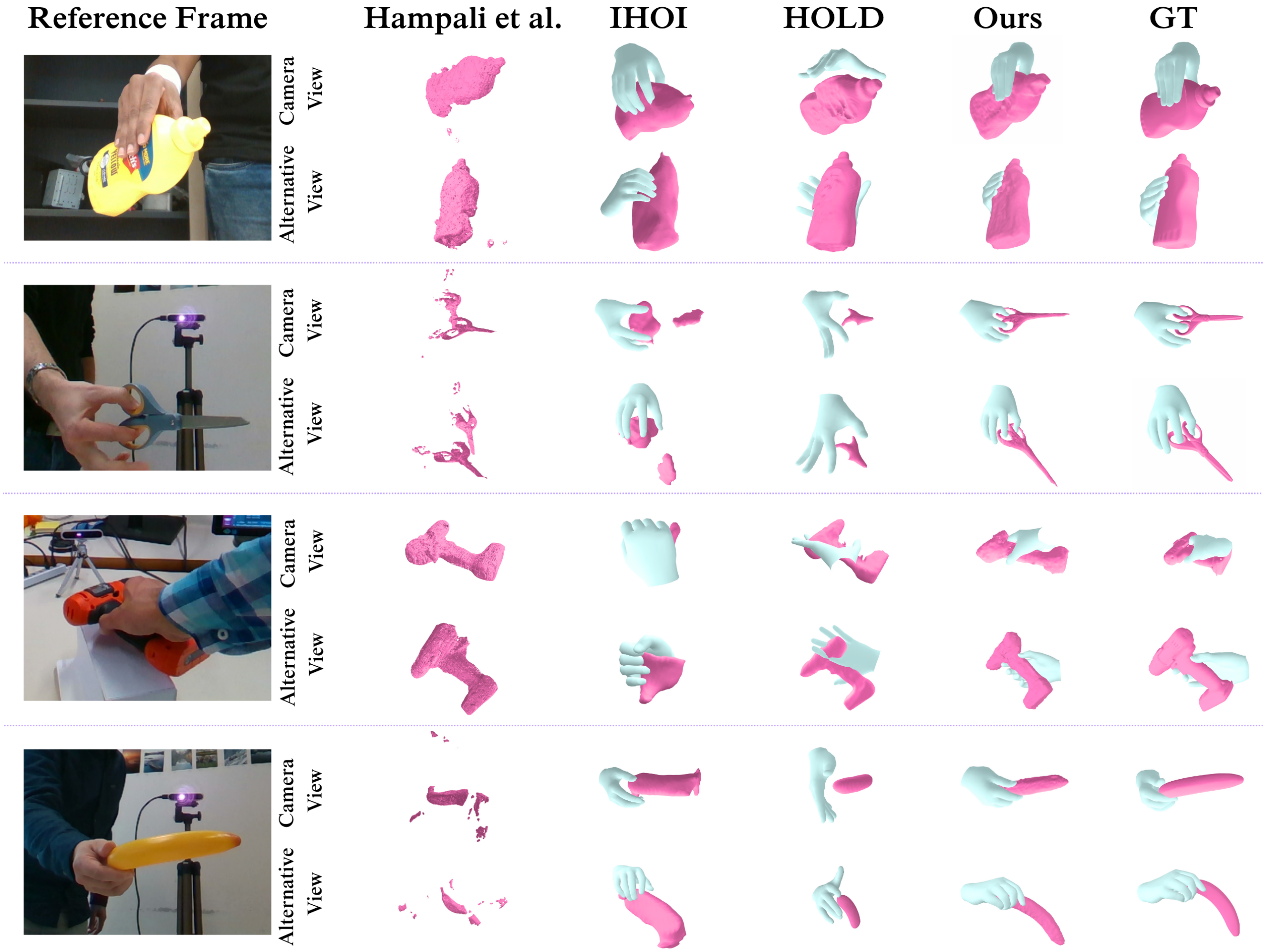

Figure 4. **Qualitative comparison with state-of-the-art methods.** Compared to prior methods, our approach achieves more accurate, stable, and geometrically consistent hand–object reconstructions across challenging scenarios. For textureless objects (banana) and thin structures (scissors), feature-based or SfM-initialized methods often produce unstable poses and distorted geometry, while ours maintains reliable in-hand poses and sharper surfaces. For heavily occluded objects (power drill, mustard bottle), prior methods frequently show misalignment or penetration, whereas AlignHOI preserves realistic contact and correct spatial relationships throughout. These results demonstrate the robustness of our align–refine strategy across diverse manipulation conditions.

4.2. Metrics and Datasets

Metrics: We evaluate our method across object reconstruction quality, hand pose accuracy, and hand–object interaction following the metrics adopted in prior works [10, 40, 45]. For object reconstruction, we apply Iterative Closest Point (ICP) alignment [2] to register the reconstructed mesh with the ground truth, and then compute the Chamfer Distance (CD) and F-score to evaluate geometric fidelity. CD measures point-to-point mesh discrepancy, while F-score (at 5 mm and 10 mm) reflects local shape accuracy after alignment. For hand pose accuracy, we use the Root-relative Mean-Per-Joint Position Error (MPJPE), measuring the mean joint distance after normalizing to the hand root. Additionally, we report the hand-relative Chamfer Distance (CD$_h$) to measure the object's shape and pose consistency relative to the hand.

Datasets: We use the HO3D-v3 dataset [12, 13] for evaluation, which contains RGB videos of hand–object interactions along with ground-truth object annotations from the YCB dataset [3]. Following the experimental setup of HOLD, we use the same set of sequences and preprocessing pipeline. To further evaluate real-world performance and generalization, we additionally collected a diverse set of in-the-wild sequences.

Table 1. **Comparison with state-of-the-art HOI reconstruction methods.** We evaluate object reconstruction accuracy, hand pose accuracy, and hand-object interaction quality. Our method achieves consistently superior performance, demonstrating significant improvements over prior approaches.

Method	CD [cm^2] ↓	F10 [%] ↑	MPJPE [mm] ↓	CD_h [cm^2] ↓
iHOI [45]	3.8	75.8	38.4	41.7
DiffHOI [46]	4.3	68.8	32.3	43.8
HOLD [10]	1.3	90.6	41.4	21.4
Ours	**0.6**	**93.4**	**4.2**	**16.4**

4.3. State-of-the-art comparison

Fig. 4 provides a qualitative comparison between our method and state-of-the-art approaches. Hampali et al. [14] represent the current in-hand object scanning paradigm, which focuses solely on reconstructing the object without modeling the hand. As the code is unavailable and only partial reconstruction sequences are publicly accessible, we follow their released sequences for comparison. IHOI [45] is a single-image method that reconstructs generic hand–object configurations without requiring category-specific 3D templates. HOLD [10] is the state-of-the-art video-based baseline that jointly reconstructs both the hand and the object, and is therefore the most relevant method to ours.

Our qualitative comparisons show that existing methods consistently struggle with some challenging cases, including textureless objects (e.g., the banana) and thin objects (e.g., the scissors). In contrast, our method produces significantly more accurate and stable reconstructions. This improvement stems from our robust in-hand object pose estimation, which is less affected by missing textures, color ambiguities, or partial occlusions. As a result, our approach maintains high reconstruction fidelity even under these challenging conditions, which is one of the key distinctions between our method and previous baselines. Moreover, for everyday manipulation scenarios involving frequent hand–object contact or occlusions, such as the power drill and mustard bottle, our method also delivers superior reconstruction quality. By incorporating both spatial and physical consistency during our hand–object interaction optimization (§ 3.4), our results exhibit realistic interaction behaviors, effectively avoiding penetrations while preserving meaningful contact throughout the entire manipulation sequence.

We conduct quantitative comparisons against both hand–object reconstruction methods and in-hand object scanning methods, as summarized in Tab. 1 and Tab. 2. Tab. 1 reports object reconstruction quality, hand pose accuracy, and hand–object interaction accuracy. Across all metrics, our method achieves the best overall performance. Tab. 2 compares our method with current in-hand object scanning methods, focusing on object reconstruction quality. We evaluate on the video sequences used by Hampali et al. [14], and the results demonstrate that our method also performs favorably in the in-hand object scanning setting.

Table 2. **Comparison with state-of-the-art in-hand object scanning methods.** All methods are evaluated on the same HO3D sequences [13] follow Hampali et al. [14] for fair comparison.

Method	CD [cm^2] ↓	F5 [%] ↑	F10 [%] ↑
Hampali et al. [14]	1.4	57.4	79.9
Jiang et al. [17]	0.6	76.0	**94.4**
Ours	**0.5**	**78.2**	**94.4**

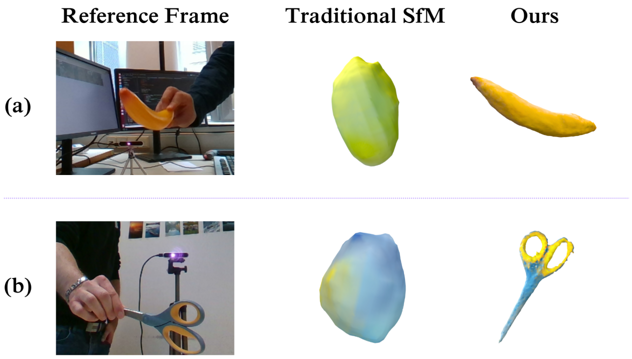

Figure 5. **Ablation study of in-hand object pose estimation.** We compare our in-hand object pose initialization with a traditional SfM-based initialization. SfM produces unstable poses when the object is textureless or partially occluded, leading to severely distorted reconstructions. In contrast, our initialization method yields stable and accurate object poses, resulting in significantly improved reconstruction quality.

4.4. Ablation Study

In-hand Object Pose Estimation vs. Traditional SfM: We study the impact of our in-hand object pose initialization by replacing it with a traditional SfM-based initialization [30, 31], while keeping all subsequent reconstruction and optimization modules unchanged. As shown in Tab. 3,

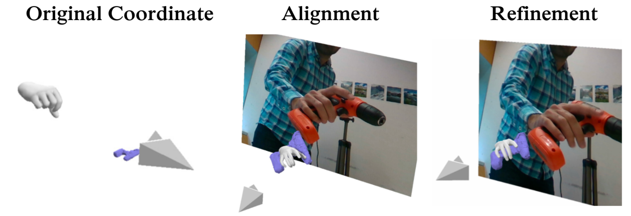

Figure 6. **Ablation study on hand–object alignment.** The hand and object are initially predicted in separate coordinate systems, causing large spatial inconsistencies. The learned 3D relative-position prior provides coarse alignment, but residual misalignment remain. With mask- and contact-based refinement, the hand and object become geometrically consistent and physically plausible. This shows that both coarse alignment and refinement are essential for stable, realistic hand–object interactions.

SfM-based initialization leads to noticeably lower reconstruction quality, especially on sequences with weak texture, repetitive patterns, or large hand-induced occlusions (see Fig. 5). Since SfM fundamentally relies on robust pixel feature tracks, it often fails to produce stable camera–object correspondences in these challenging settings, resulting in drifting or inconsistent object poses.

In contrast, our initialization strategy is designed specifically for the HOI scenario. By leveraging feature-based template matching and geometric-consistency verification, it produces stable and accurate pose estimates even when visual features are scarce or occluded. These reliable initial poses provide a much stronger starting point for the subsequent reconstruction pipeline, ultimately yielding a significantly more consistent and faithful object reconstruction. This comparison highlights the limitations of feature-dependent SfM under HOI conditions, and demonstrates the necessity of our tailored in-hand pose estimation approach.

Effect of Object Pose Refinement: We further evaluate the importance of our object pose refinement module by disabling it after the object pose matching. In this setting, we select the top five template candidates, identify the best one, and directly use its template pose without any refinement. As shown in Tab. 3, skipping the refinement step leads to degradation in object reconstruction quality: the lack of feature-based alignment and iterative optimization results in residual pose errors that propagate into the reconstructed geometry. In contrast, our refinement stage leverages DINOv2 feature correspondences and reprojection-error minimization to correct these inaccuracies, producing poses that are geometrically consistent with the input image. This yields better object reconstruction results. Overall, this ablation highlights that coarse template matching alone is insufficient, and precise per-frame refinement is critical for achieving high-fidelity and temporally stable reconstructions.

Table 3. **Ablation study.** We analyze the impact of two key components of our method: (1) replacing our in-hand object pose initialization with a traditional SfM-based initialization, and (2) disabling the object pose refinement module and directly using the retrieved discrete template pose.

	CD $[\mathrm{cm}^2]\downarrow$	F5 [%] ↑	F10 [%] ↑
SfM-based pose	5.8	60.4	75.2
w/o refinement	1.1	71.8	87.8
Ours	**0.6**	**76.4**	**93.4**

Effect of Hand–Object Alignment: As shown in Fig. 6, the hand and object are initially predicted in different coordinate systems, resulting in large spatial inconsistencies and unrealistic interaction. Introducing our learned 3D relative-position prior provides a coarse but meaningful alignment, bringing the hand and object into the correct spatial relationship. However, without further refinement, residual discrepancies remain, leading to inaccurate contact and occasional penetration. With the full refinement module, which enforces mask consistency and contact while suppressing penetration, the hand and object become geometrically compatible and interact in a physically plausible manner. This ablation demonstrates that both the coarse alignment and the refinement stage are essential for producing stable and realistic hand–object interactions.

5. Conclusion

In this paper, we introduced AlignHOI, a new framework for reconstructing dynamic hand–object interactions from monocular video via alignment and refinement. Our method converts the ill-posed problem of category-agnostic in-hand object pose estimation into a tractable finite template-matching task, enabling stable coarse alignment even under severe occlusions, low-textured surfaces, and thin-object geometries. The retrieved poses are further refined via geometric consistency. Building on these robust poses, we reconstruct the object using an efficient implicit neural representation, which enables fast and high-fidelity surface recovery. To resolve cross-coordinate inconsistencies between the hand and object, we introduce a physically and spatially grounded optimization strategy that uses a learned 3D relative-position prior along with contact and penetration constraints. This refinement yields coherent, realistic, and stable 3D interactions. Extensive experiments demonstrate that AlignHOI not only achieves state-of-the-art reconstruction accuracy but also delivers significantly improved hand–object alignment and interaction plausibility.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7061–7071, 2025. 2

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 6

[3] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 6

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12417–12426, 2021. 2, 3

[5] Jiayi Chen, Mi Yan, Jiazhao Zhang, Yinzhen Xu, Xiaolong Li, Yijia Weng, Li Yi, Shuran Song, and He Wang. Tracking and reconstructing hand object interactions from point cloud sequences in the wild. In *Proceedings of the AAAI conference on artificial intelligence*, pages 304–312, 2023. 3

[6] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European conference on computer vision*, pages 231–248. Springer, 2022. 3

[7] Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. Hort: Monocular hand-held objects reconstruction with transformers. *arXiv preprint arXiv:2503.21313*, 2025. 3, 5, 2

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 3

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 2

[10] 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*, pages 494–504, 2024. 2, 3, 5, 6, 7

[11] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2

[12] 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*, pages 3196–3206, 2020. 6

[13] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11090–11100, 2022. 6, 7

[14] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17079–17088, 2023. 3, 7

[15] 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*, pages 11807–11816, 2019. 3

[16] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 3

[17] Shijian Jiang, Qi Ye, Rengan Xie, Yuchi Huo, and Jiming Chen. Hand-held object reconstruction from rgb video with dynamic interaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12220–12230, 2025. 2, 3, 5, 7

[18] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[19] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000. 2

[20] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944. 2

[21] 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*, pages 7037–7047, 2025. 3, 5

[22] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 2

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2

[24] 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. 5, 2

[25] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pages 163–182. Springer, 2024. 4, 1

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[27] 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*, pages 12242–12254, 2025. 4, 5

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[29] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 2, 3, 4

[30] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 7

[31] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 7

[32] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[33] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3352–3360. IEEE, 2025. 2

[34] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 4

[35] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2

[36] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 3

[37] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering, 2022. 2, 3

[38] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4): 376–380, 2002. 4

[39] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[40] 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. *arXiv preprint arXiv:2508.05506*, 2025. 2, 3, 6

[41] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023. 2, 5

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023. 3

[43] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025. 2, 4, 1

[44] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2750–2760, 2022. 3

[45] 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*, pages 3895–3905, 2022. 6, 7

[46] 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*, pages 19717–19728, 2023. 3, 7

[47] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, et al. Human performance modeling and rendering via neural animated mesh. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 2

AlignHOI: Hand–Object Reconstruction via Alignment and Refinement

Supplementary Material

This supplementary material provides extended details and results that complement the main paper. Section A presents further explanations and derivations of our method, while Section B includes additional experiment details, ablation studies, and quantitative results that extend the analysis in the main paper.

A. More Details on Our Method

This section provides additional derivations, algorithmic explanations, and implementation details for our method introduced in Section 3 of the main paper.

A.1. Preparation: Coarse Object Prior

We use Amodal3R [43] to generate the initial coarse object mesh $\hat{\mathcal{M}}_o$ for rendering template views. Since our input is a video rather than a set of manually selected views, we introduce a fixed view-selection strategy to construct the four-view input following the default multi-view setting of Amodal3R. This deterministic procedure ensures reproducible view inputs and improves the stability of the generated results. Specifically, we uniformly divide the video into four temporal segments $\{\mathcal{S}_k\}_{k=1}^4$ and, from each segment, select the frame with the largest unoccluded object area. Let $\mathcal{A}(t)$ denote the number of visible object pixels in frame t. Using this visibility measure, we identify one representative frame from each temporal segment and then sort the four selected frames in descending order of visibility, such that higher-quality views appear earlier in the input sequence:

$$\begin{aligned} t_k^\star &= \arg\max_{t \in \mathcal{S}_k} \mathcal{A}(t), \quad k = 1, \dots, 4, \\ \text{s.t.} \quad &\mathcal{A}(t_1^\star) \geq \mathcal{A}(t_2^\star) \geq \mathcal{A}(t_3^\star) \geq \mathcal{A}(t_4^\star). \end{aligned}$$

Our view-selection strategy preserves temporal diversity, maximizes object visibility in each segment, and prioritizes the most informative views, resulting in more reliable generation quality. As shown in Fig. A, we compare four view-selection strategies for generating the coarse object prior from a video. Our method, which selects one high-visibility frame per temporal segment and orders them by visibility, produces the most stable and complete meshes. Using the same per-segment frames but keeping their temporal order yields slightly inferior results, showing mild distortions due to suboptimal view ordering. Randomly sampling four frames from the entire video leads to highly unstable outputs, with breakages caused by occluded or uninformative views. Using only the single frame with the largest visible area can produce plausible mesh but lacks viewpoint diversity, and thus fails to generalize consistently to custom or complex object. These results highlight the importance of both temporal coverage and visibility-aware ordering for reliable coarse object generation.

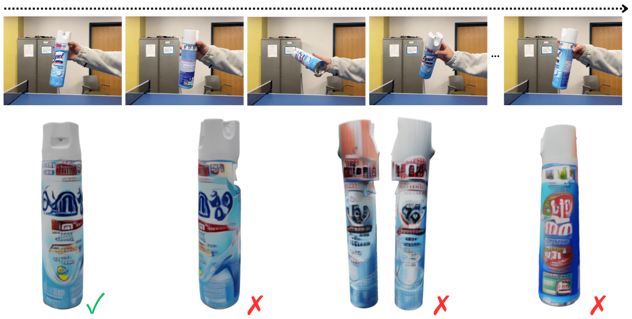

Figure A. **Our visibility-aware view-selection strategy produces stable and complete coarse objects.** Given a video input, we compare different view-selection strategies for generating the coarse object prior. Our method, which selects one high-visibility frame per temporal segment and orders them by visibility, yields consistent meshes, whereas temporal-order, random-view, and single-view inputs produce inferior results.

During inference, we fix the sampler parameters as $\lambda_{\text{geo}} = 9.5$ and $\lambda_{\text{tex}} = 5.0$. Generally, λ_{geo} controls the guidance strength for geometry generation, while λ_{tex} controls the guidance strength for texture generation.

A.2. Object Pose Estimation

Object Pose Matching: To obtain the object template set, following [25], we render the 3D object model from multiple viewpoints. We uniformly sample 57 viewing directions over the viewsphere and apply 14 evenly spaced in-plane rotations for each direction, resulting in a total of 798 templates per object. We extract DINOv2 patch features from each template and reduce them to 256 dimensions via PCA. All projected features are clustered using k-means into a vocabulary of 2,048 visual words. Each template is then represented by a 2,048-dimensional descriptor.

Given a query image, we extract DINOv2 features on a 14×14 grid inside the object mask and project them into the same PCA space. The descriptor for the query image is then computed using the shared visual vocabulary. We compute cosine similarities $s(I, T_i)$ and retrieve the top-5 most similar templates. Finally, we perform a geometric consistency check to obtain the final template selection.

Object Pose Refinement: Given the coarse pose initialization provided by the retrieved template pose $\mathbf{p}_{\text{win}}^{\text{temp}} \in \mathcal{P}$, we refine the pose using Levenberg–Marquardt [20, 22] to iteratively minimize the reprojection error over the established 2D–3D correspondences. Formally, we solve

$$\theta^* = \arg\min_{\theta} \sum_i \left\| \pi_\theta(X_i) - x_i \right\|^2, \tag{7}$$

where X_i are 3D object points in the model coordinate frame, x_i are the corresponding 2D image locations, and π_θ is the camera projection function with pose parameters $\theta = (R, t)$ representing the model-to-camera transformation. With camera intrinsics K, the projection becomes

$$\pi_\theta(X) = K[R \mid t]X, \tag{8}$$

where X denotes a homogeneous 3D point. Starting from $\mathbf{p}_{\text{win}}^{\text{temp}}$, the optimization iteratively updates (R, t) until convergence, yielding the continuous pose estimate $\mathbf{P}_{\text{obj}}$ for the current frame.

Discussion of DINOv2-Based Feature Choices: DINOv2 provides dense, highly discriminative patch-level features that are well suited for our template-based object pose estimation pipeline. These features exhibit strong robustness to the domain gap between rendered templates and real images, where differences in lighting, material appearance, and sensor noise can otherwise degrade matching performance. Moreover, DINOv2 features support efficient large-scale retrieval: once extracted, they are projected into a compact PCA space and indexed for fast nearest-neighbor search. In addition, the features produce stable local correspondences that benefit geometric consistency checking and pose refinement, enabling reliable matching even for texture-poor or partially occluded objects.

In our experiments, we explored multiple variants of DINO-based representations, including different model sizes and feature-extraction layers from both DINOv2 [24] and DINOv3 [32]. A general observation is that larger models tend to produce stronger and more distinctive features when an appropriate intermediate layer is selected. Among the configurations we tested, both DINOv2 ViT-g/14 (layer 30) and DINOv3 ViT-7B/16 (layer 38) yielded competitive performance. However, DINOv3 ViT-7B/16 operates at a coarser spatial granularity (16×16 patches), incurs several times higher computational cost, and does not provide clear benefits for our object-level matching task. Considering the trade-off between accuracy, feature granularity, and inference efficiency, we adopt DINOv2 ViT-g/14 features from the 30th layer as our default representation.

A.3. Efficient Implicit Object Reconstruction

We provide here a brief technical clarification of how we achieve efficient implicit object reconstruction in our HOI pipeline. We aim to accelerate implicit object reconstruction, as existing HOI reconstruction pipelines typically require dozens of hours to optimize a single sequence [10, 17], making large-scale reconstruction and subsequent development impractical.

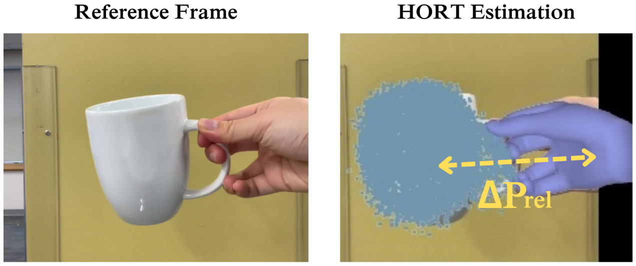

Figure B. **Visualization of the relative hand–object displacement predicted by HORT [7].** Given a reference frame (left), we use HORT to estimate the relative translation $\Delta\mathbf{p}_{\text{rel}}$ between the hand and the object in its coordinate system (right).

Although a naïve combination of neural surface methods [39] with Instant-NGP [23] appears appealing for fast optimization, it does not work in practice. Instant-NGP does not support efficient second-order derivative computation in backpropagation, which is required for enforcing the Eikonal constraint in neural surface formulations. Instant-NSR [47] addresses this limitation using finite-difference approximations of second-order derivatives, but such approximations introduce numerical inaccuracies and often lead to unstable training, as discussed in NeuS2 [41]. NeuS2 provides a more principled and efficient alternative by introducing an analytic and stable formulation of the required second-order derivatives. We integrate NeuS2 into the dynamic HOI pipeline and observe both substantial speedup and improved reconstruction accuracy under challenging hand–object interactions. This integration enables fast, stable, and high-fidelity reconstruction, greatly enhancing the practicality of our system by supporting efficient experiments and future scaling to larger datasets and more complex HOI scenarios.

A.4. Hand-Object Interaction Optimization

Hand-Object Alignment: We present the detailed mathematical formulation of the hand–object alignment procedure here. The object position in the target coordinate system is obtained by transforming the canonical mesh center:

$$\mathbf{p}_{\text{obj}}^{\text{target}} = s_{\text{scene}} \cdot \Big(R \cdot \big(s_{\text{obj}} \cdot \text{mean}(\text{denorm}(V_{\text{obj}}^{3d})) \big) + t \Big), \tag{9}$$

where s_{scene} is the global scene scaling factor, R and t denote the rigid transformation applied to the object, s_{obj} is the object-specific scaling factor, and V_{obj}^{3d} is the set of object vertices.

As illustrated in Fig. B, HORT [7] provides the relative displacement between the hand and the object in its own coordinate system:

$$\Delta \mathbf{p}_{\text{rel}} = \mathbf{p}_{\text{hand}}^{\text{HORT}} - \mathbf{p}_{\text{obj}}^{\text{HORT}}, \tag{10}$$

where $\mathbf{p}_{\text{hand}}^{\text{HORT}}$ and $\mathbf{p}_{\text{obj}}^{\text{HORT}}$ denote the hand and object centroids, respectively.

Since the relative displacement is invariant across coordinate systems, the hand position in the target space is computed as

$$\mathbf{p}_{\text{hand}}^{\text{target}} = \mathbf{p}_{\text{obj}}^{\text{target}} + \Delta \mathbf{p}_{\text{rel}}. \tag{11}$$

This derivation ensures that the hand and object are consistently aligned in the unified coordinate system by preserving their relative configuration. Note that this alignment only accounts for translation, as it is computed from the point cloud centroids. Subsequent hand pose refinement further optimizes the remaining degrees of freedom, including rotation and articulation details.

Hand-Object Interaction Refinement: We provide detailed definitions of the losses used in the hand parameters refinement stage.

To ensure silhouette consistency, we define the masked IoU loss as:

$$\begin{aligned} \mathcal{L}_{\text{mask}} &= 1 - \frac{\text{Int}}{\text{Uni}}, \\ \text{Int} &= \sum_i V_i\, M_{\text{pred},i} M_{\text{gt},i}, \\ \text{Uni} &= \sum_i V_i\, M_{\text{pred},i} + \sum_i V_i\, M_{\text{gt},i} - \text{Int}, \end{aligned} \tag{12}$$

where V_i denotes whether pixel i is valid:

$$V_i = \begin{cases} 1, & \text{pixel } i \text{ is valid for supervision,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

This validity mask excludes pixels belonging to the other class, ensuring, for example, that object regions do not affect the hand-mask loss. Unlike simple pixel-wise overlap losses that may fail to penalize completely disjoint masks, the IoU formulation provides a more stable gradient when the overlap is small or zero. We also adopt an occlusion-aware computation. During hand mask fitting, pixels belonging to the object mask are ignored in the loss to prevent erroneous penalization in occluded regions.

We encourage the hand to touch the object by minimizing the distance from hand contact vertices to the nearest object vertices:

$$\mathcal{L}_{\text{contact}} = \frac{1}{|V_{\text{h}}^c|} \sum_{v \in V_{\text{h}}^c} \min_{u \in V_{\text{o}}} \|v - u\|_2^2. \tag{14}$$

where V_{h}^c denotes the hand contact-region vertices, and V_{o} is the set of object-surface vertices.

Table A. HO3D sequences used for experiments.

Sequence name	**Object**	**Total Frames**
ABF12	bleach	222
ABF14	bleach	222
BB12	banana	187
BB13	banana	257
GPMF12	potted meat	184
GPMF14	potted meat	175
GSF12	scissors	167
GSF13	scissors	255
MC1	cracker box	144
MC4	cracker box	144
MDF12	power drill	449
MDF14	power drill	449
ShSu10	sugar box	296
ShSu12	sugar box	296
SM2	mustard	144
SM4	mustard	144
SMu1	mug	287
SMu40	mug	320

To discourage hand–object intersections, we use an SDF-based penetration loss:

$$\mathcal{L}_{\text{pene}} = \frac{1}{|V_{\text{h}}|} \sum_{v \in V_{\text{h}}} \max(0, -\Phi_{\text{o}}(v)), \tag{15}$$

where Φ_{o} is the object signed distance field and V_{h} is the set of hand vertices.

Overall, the refinement loss is expressed as

$$\mathcal{L}_{\text{ref}} = \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{contact}} \mathcal{L}_{\text{contact}} + \lambda_{\text{pene}} \mathcal{L}_{\text{pene}}, \tag{16}$$

where λ_{mask}, λ_{contact}, and λ_{pene} are the respective weights.

B. More Experiments

B.1. More Experiment Details

In Tab. A, we list the HO3D sequences used for experiments, following the same evaluation protocol as HOLD. These sequences cover diverse hand–object configurations and interaction patterns, ensuring a comprehensive assessment of reconstruction performance.

In Fig. C, we also provide the corresponding coarse object priors generated for each sequence. We use these meshes as geometric priors during the object pose matching stage, where they supply 3D-consistent renderings for template retrieval. Although the generated priors exhibit noticeable differences in both geometry and texture, our method remains robust and consistently produces high-quality reconstruction results.

Table B. **Ablation study of object pose estimation.** We evaluate the contribution of each component in our pipeline. Top-n retrieval, the template pose prior, and refinement all play essential roles in achieving stable and accurate in-hand object pose estimation.

Method	Top-n	Template Pose	Refinement	CD [cm^2] ↓	F5 [%] ↑	F10 [%] ↑	ATE [mm] ↓
SfM Init	✗	✗	✗	5.8	60.4	75.2	4.1
w/o Top-n (Top-1 only)	✗	✓	✓	0.7	74.2	92.1	3.2
w/o Template Pose	✓	✗	✓	1.0	71.3	89.6	3.5
w/o Refinement	✓	✓	✗	1.1	71.8	87.8	3.5
Top-3 candidates	✓	✓	✓	**0.6**	76.2	92.5	**3.0**
Top-10 candidates	✓	✓	✓	0.7	76.0	91.9	**3.0**
Ours (full, Top-5)	✓	✓	✓	**0.6**	**76.4**	**93.4**	**3.0**

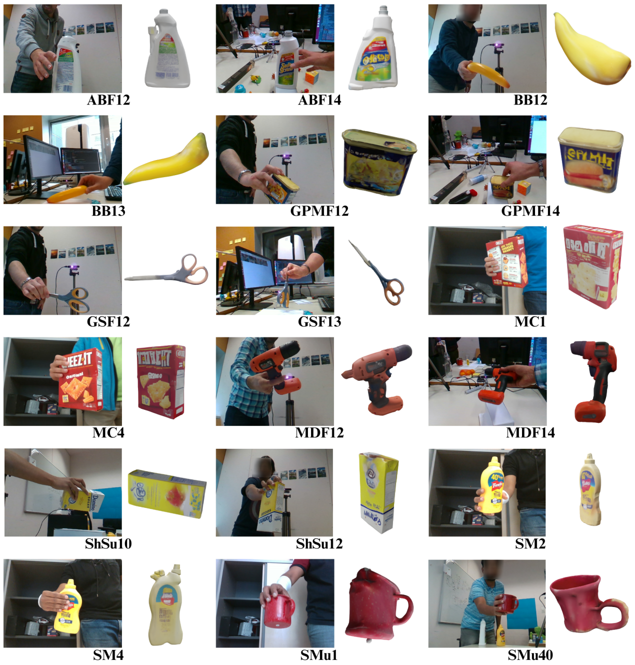

Figure C. **Coarse object priors generated for each sequence.** For each sequence, the left image shows the reference frame and the right shows the object prior. Despite variations in geometry and texture, these priors provide sufficient 3D cues for robust template-based pose matching.

B.2. More Ablation Studies

Tab. B provides comprehensive ablation results to further demonstrate the contribution of each component in our object pose estimation pipeline. All experiments are evaluated following the metrics described in Section 4.2 of the main paper, including CD, F5, and F10. In addition, we follow the SLAM literature and introduce the Absolute Trajectory Error (ATE) to further evaluate pose accuracy [34], where we first align the predicted trajectory with the ground-truth trajectory using Umeyama alignment [38] and then compute the trajectory error in millimeters.

SfM Initialization: We replace our template-based pose estimation pipeline with a traditional SfM pipeline. SfM performs poorly in HOI videos due to severe hand–object occlusions, rapid object motion, and the lack of stable texture features on many objects. This leads to significantly degraded CD, F5, F10, and ATE metrics.

w/o Top-n (Top-1 Only): We disable Top-n template retrieval and directly select the template with the highest similarity scores, followed by pose refinement. Although DINOv2 features are robust, cosine similarity does not necessarily reflect geometric consistency, causing the top-1 template to often correspond to an incorrect pose. The subsequent refinement step can alleviate small deviations but cannot correct a fundamentally wrong template choice, highlighting the importance of the Top-n candidate set and geometric verification.

w/o Template Pose: We disable the discrete template pose and estimate the pose solely using patch correspondences via PnP optimization. While correspondences and PnP can still recover a reasonable pose, the optimization becomes more vulnerable to local mismatches and partial occlusions without the template pose prior. This results in degraded accuracy, demonstrating that the template pose provides an essential coarse geometric prior rather than merely supplying correspondences.

w/o Refinement: We remove the final LM-based pose refinement and directly use the discrete template pose as the output. Since template poses belong to a discrete pose space and cannot perfectly match the input image, removing refinement retains the discretization error, causing a noticeable drop across all metrics. This highlights the importance of continuous refinement for accurate pose estimation.

Top-n Candidate Set Size: We evaluate Top-3, Top-5, and Top-10 and observe that all configurations outperform Top-1 by a clear margin. This confirms that, compared to selecting only the Top-1 template, constructing a candidate set and applying geometric consistency filtering consistently improves performance. Among these settings, Top-5 achieves the best overall accuracy, and we therefore adopt Top-5 as the default configuration in our full pipeline.

Summary: Across all settings, our ablations validate the importance of each module in our object pose estimation pipeline. Top-n retrieval and geometric consistency ensure reliable template selection, the template pose provides a strong coarse prior, and refinement is critical for precise optimization. Together, these components enable robust pose estimation even under occlusion and low-texture conditions.

B.3. More Qualitative Results

Continuous Hand–Object Interaction Sequences. To further demonstrate the stability and temporal consistency of our pipeline, we visualize complete hand–object interaction sequences in Fig. D. Our method produces smooth, temporally coherent trajectories for both the hand and object, enabling physically plausible interaction modeling over long video spans. The reconstructed geometry and poses remain consistent across frames despite severe self-occlusion, rapid motion, and viewpoint changes. These results illustrate that our system is not limited to isolated frames, but can robustly support continuous HOI reasoning, analysis, and downstream applications such as interaction understanding and dynamic scene reconstruction.

In-the-wild Results: As noted in the main paper, we additionally collect a diverse set of in-the-wild sequences to evaluate the real-world performance and generalization ability of our method. In Fig. E, we present several challenging cases to stress-test the robustness and effectiveness of our method. These include: (1) a Grogu doll with complex wrinkles and geometry; (2) a cleaning spray bottle with thin-sheet partial structures; (3) a texture-less pink bottle with a slender shape; (4) a toy water gun with rich geometric details; (5) and a white mug that is texture-less and reflects colored environmental lighting. These cases highlight our method's ability to handle diverse object types, varying material properties, and imperfect generated coarse priors while still producing stable and accurate reconstructions.

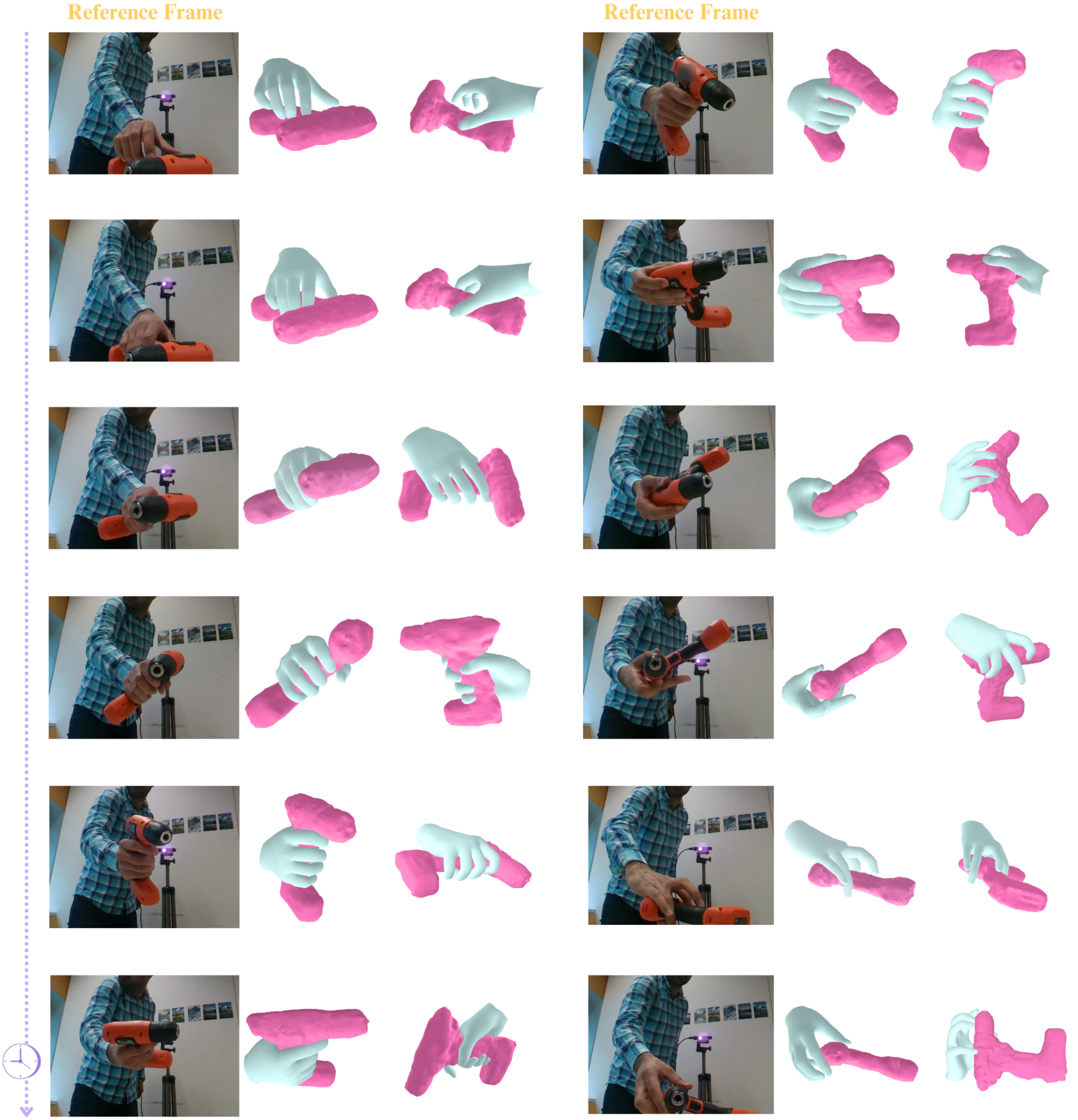

Figure D. **Continuous Hand–Object Interaction Sequences.** Given a sequence in HO3D, we show the temporally ordered HOI reconstruction results. Each timestep begins with the reference frame, followed by our reconstructed results in the camera view and an alternative view. Across time, our method maintains consistent geometry, stable hand–object contact, and smooth motion, demonstrating robust and temporally coherent HOI reconstruction even under self-occlusion and rapid motion.

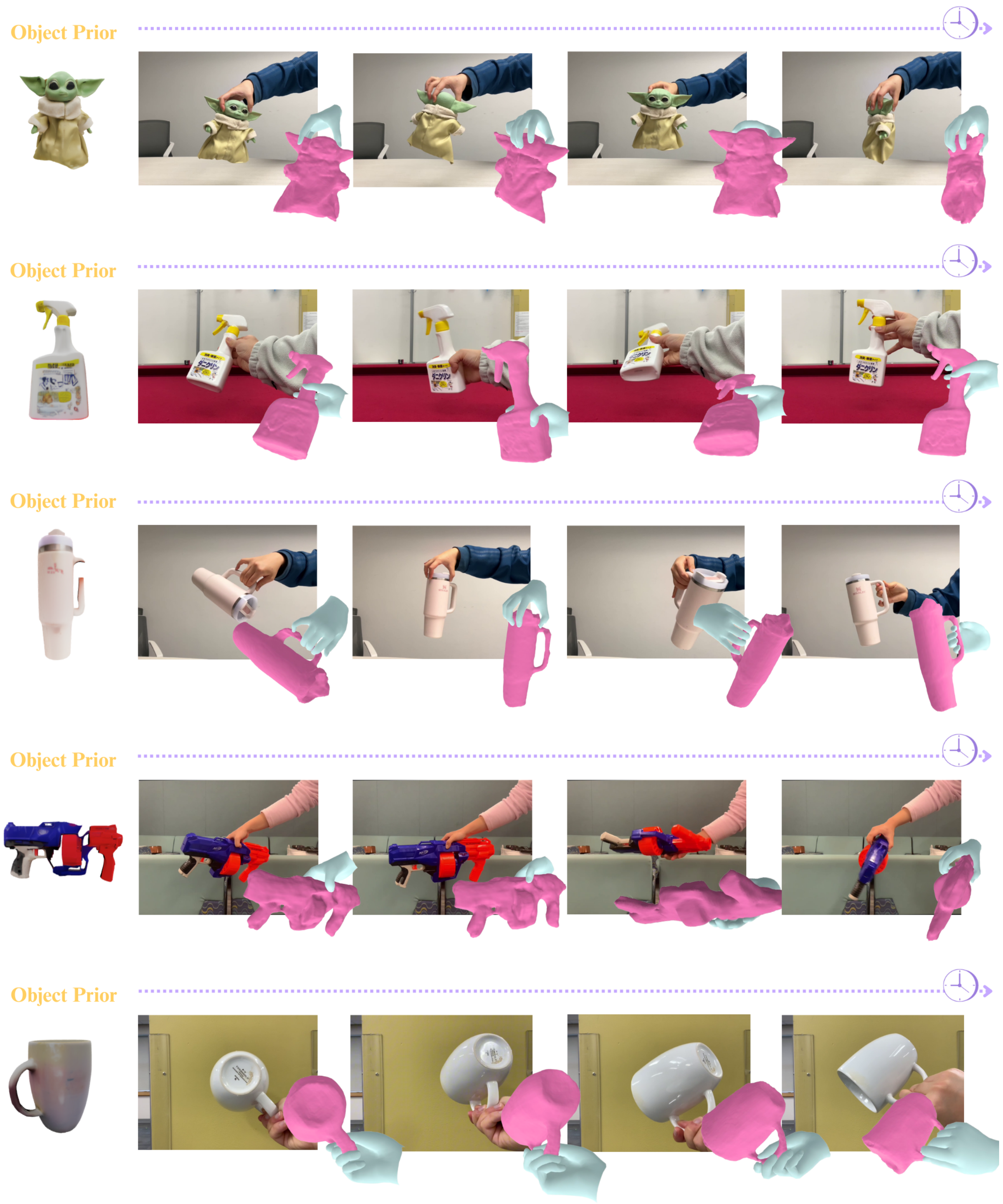

Figure E. **In-the-wild results.** For each example, the leftmost image shows the generated object prior, and the remaining images show HOI reconstruction results across different frames. These diverse in-the-wild cases demonstrate that our method produces stable and accurate reconstructions under varying object types, shapes, materials, and lighting conditions.