arXiv:2512.07394v2 [cs.CV] 2 Jun 2026

Reconstructing Objects along Hand Interaction Timelines in Egocentric Video

Zhifan Zhu[1] Siddhant Bansal[1] Shashank Tripathi[2] Dima Damen[1]
[1]University of Bristol, UK [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
https://zhifanzhu.github.io/objects-along-hit

Abstract

We introduce the task of Reconstructing Objects along Hand Interaction Timelines (ROHIT). We first define the Hand Interaction Timeline (HIT) from a rigid object's perspective. In a HIT, an object is first static relative to the scene, then is held in hand following contact, where its pose changes. This is usually followed by a firm grip during use, before it is released to be static again w.r.t. to the scene. We model these pose constraints over the HIT, and propose to propagate the object's pose along the HIT enabling superior reconstruction using our proposed Constrained Optimisation and Propagation (COP) framework. Importantly, we focus on timelines with stable grasps – i.e. where the hand is stably holding an object, effectively maintaining constant contact during use. This allows us to efficiently annotate, study, and evaluate object reconstruction in videos without 3D ground truth.

We evaluate our proposed task, ROHIT, over two egocentric datasets, HOT3D and in-the-wild EPIC-Kitchens. In HOT3D, we curate 1.2K clips of stable grasps. In EPIC-Kitchens, we annotate 2.4K clips of stable grasps including 390 object instances across 9 categories from videos of daily interactions in 141 environments. Without 3D ground truth, we utilise 2D projection error to assess the reconstruction. Quantitatively, COP improves stable grasp reconstruction by $6.2 - 11.3\%$ *and HIT reconstruction by up to* 24.5% *with constrained pose propagation.*

1. Introduction

Accurately reconstructing three-dimensional hand-object interactions is key to unlocking many perception problems, including fine-grained understanding of interactions, but also potential applications in augmented reality, robotic imitation learning and human-machine interactions.

Efforts to reconstruct the object, in-hand, have typically used 3D ground-truth supervision, paired with 3D [4, 61, 78] or 2D input [10, 29, 35, 70, 73]. While such approaches are typically developed to be view-independent, early insights [6, 29, 49] demonstrate that egocentric footage remains significantly challenging, partly due to the considerable occlusion of the object by the hand.

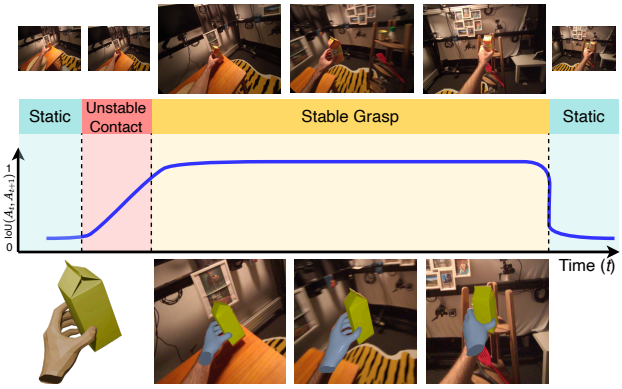

Figure 1. Sample HIT sequence from HOT3D [3] with reconstruction results by our method. We illustrate the three types of temporal segments in hand-object interactions: Static: where the object is static relative to the scene, Unstable Contact: where the hand is firming its grip on the object; and Stable Grasp: where hand is securely holding the object stably, until it is Static again when put down. The plot illustrates the IoU of the in-contact vertices across neighbour frames; for formal definition, refer to Sec. 3.1.

In this work, we particularly focus on egocentric videos, and **for the first time** consider the complete timeline of the interaction – before contact, when in-hand, and after release. We model interactions with *functional intent* – i.e. where the object is operated or moved securely, rather than simply poked or touched. We show that during these functional grasps, the same hand and object vertices remain in contact but the object's pose still changes relative to the hand due to finger and hand articulations.

To estimate the object pose along this timeline, we propose constraints that depend on the type of interaction segment (Static, Unstable Contact or Stable Grasp) as shown in Fig. 1. We propose the Constrained Optimisation and Propagation (COP) framework to optimise the object pose during each interaction segment, then propagate the pose to initialise the next segment. We show sample reconstructions

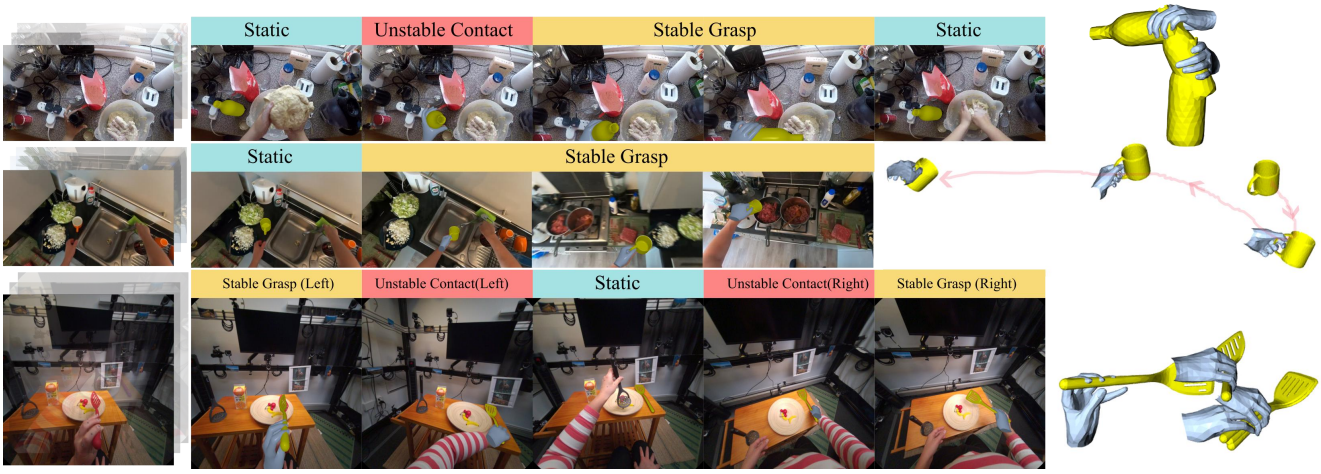

Figure 2. **Qualitative results.** Given a Hand Interaction Timeline (HIT) - with an object in Static, Unstable Contact and Stable Grasp interaction segments, our method, COP, reconstructs hand (blue) and object (yellow) meshes along the HIT. We show input frames (left), projected meshes (middle) and meshes in 3D world coordinate system (right). Rows 1-2 from EPIC-HIT and row 3 from HOT3D-HIT.

along the HIT in Fig. 2.

To summarise, our contributions are as follows:

- We propose the task of Reconstructing Objects along Hand Interaction Timelines (**ROHIT**) to reconstruct 3D poses, for rigid known objects, in egocentric videos.
- We propose the Constrained Optimisation and Propagation (**COP**) framework that optimises the object pose for various interaction segment types.
- We curate hand interaction timelines from the egocentric HOT3D dataset, to evaluate our method with 3D ground truth. We refer to this as **HOT3D-HIT** dataset.
- We label $2.4K$ stable grasps clips from EPIC-Kitchens, along with 96 HIT, which we call the **EPIC-HIT** dataset.
- We evaluate COP on both datasets and show $6.2 - 11.3\%$ improvement in stable grasp reconstruction and up to 24.5% gain in HIT reconstruction with propagation.

2. Related Works

Here we discuss work on hand pose estimation, object pose estimation and joint hand-object reconstruction. For comparison of hand object reconstruction datasets, see Tab. 1.

3D Hand Pose Estimation. Estimating 3D hand pose from images has been proposed for both free hands and hands in-interactions. FrankMocap [56] is a commonly used CNN-based model in many hand-object reconstruction methods [6, 29, 49, 73, 74]. METRO [40] proposes to use a transformer on top of the CNN feature for regression. WildHands [54] addresses the perspective distortion for egocentric hand pose estimation. Recently, fully transformer-based methods [19, 39, 50, 52, 80] train on scaled training data with higher network capacity, they achieve superior performance. The transformer based HaMeR [50] has been recently used as guidance for full body estimation task [76] to show its usefulness. In this work, our in-the-wild pipeline also uses HaMeR [50] due to its robust performance.

3D Object Pose Estimation. A full review of works on estimating 3D objects pose from single image is out of our scope. Here, we focus on relevance for hand-object interaction scenarios. To estimate 3D object pose, several works [8, 25, 45, 48, 64] assume a known object shape template and estimate 6-DoF pose by fitting it to 2D image features (e.g. masks or keypoints), using either 2D-3D [8, 48] or 3D-3D correspondences [25, 45, 64]. Other methods [24, 69] estimate object pose and shape jointly, however, at the cost of geometric fidelity. While effective, these works typically assume unoccluded objects.

3D Hand-Object Reconstruction. Methods are grouped into two categories. The first category, known-CAD methods, assumes that object CAD models are given and fits 3D shapes into 2D observations. These can further be classified into data-driven [1, 12, 33, 38, 41, 42, 63, 66, 70, 71] or optimisation-based [6, 29, 49] methods. Data-driven methods learn to jointly reconstruct hands and objects from seen object examples, whereas optimisation-based methods address the reconstruction by directly fitting to 2D signals. RHO [6] is the first optimisation based single-frame method. The optimisation-based methods [6, 29, 49] share the same pipeline where hand/object is first independently optimised, followed by joint optimisation with physical constraint terms. HOMan [29] generalises to multiple frames and includes temporal smoothness of mesh vertices over time.

The second category, CAD-agnostic methods, estimates the object pose without using explicit CAD models. Many CAD-agnostic methods [10, 11, 28, 35, 53, 73, 75, 77] learn object shape priors, or retrieve from (generative-)object pools [2, 34, 44]. The other line of methods [21, 27, 32, 67]

uses neural networks to fit the underlying object shape from multiple views. We experimentally show that CAD-agnostic methods are incapable of generalising to in-the-wild egocentric videos.

Our method belongs to the first category, adopting a simplified assumption needed for the challenges of in-the-wild reconstruction. This also allows us to focus on HIT reconstruction, which has not been attempted before. Different from all previous optimisation-based methods, we examine the object's relative motion through various segments and reconstruct it throughout the HIT.

3. The ROHIT Task

3.1. What is a Hand Interaction Timeline?

A Hand Interaction Timeline (HIT) for a given object is a video sequence that focuses on the hand interacting with one object to either use or move that object. From the object's perspective, the HIT can be divided into many contiguous segments of three types:

- Static: before/after a hand interaction, the object is typically supported by a surface and is thus static relative to the scene/world.
- Stable Grasp: the hand grasps the object firmly, allowing functional usage and secure movement (details below).
- Unstable Contact: the object is neither Static nor in Stable Grasp, *e.g.* when a grip is being formed on the object.

As in Fig 2, the HIT can involve many temporal segments (e.g. Static → Unstable Contact → Stable Grasp → Unstable Contact → Stable Grasp → Static). Our task is to reconstruct the hand-object interaction along the full HIT.

The term Stable Grasp has been previously used in human grasp analysis [5, 15, 22]. While definitions vary, they centre around the object being "held securely with one hand, irrespective of the hand orientation" [22]. Intuitively, this means that *the same hand and object vertices remain in contact for the grasp duration*.

Formally, for any pair of frames i and j within the Stable Grasp, we use S_i and S_j to denote the in-contact area on the object surface, and intersection-over-union $\text{IOU}(S_i, S_j)$ between in-contact areas. Following above intuition, the duration of the stable grasp is defined as:

$$\begin{aligned} [l^*, r^*] &= \underset{l,r}{\text{argmax}}\,(r - l) \\ \text{s.t. } \text{IOU}(S_i, S_j) &> \tau \qquad \forall l \leq i < j \leq r \end{aligned} \tag{1}$$

where τ specifies the minimum IOU threshold. The $\text{argmax}\,(r - l)$ implies the duration of the stable grasp. Importantly, while there is a *consistent contact area*, the hand orientation and finger articulations/poses vary during the stable grasp. This allows the object pose to vary relative to the hand – see Fig. 3.

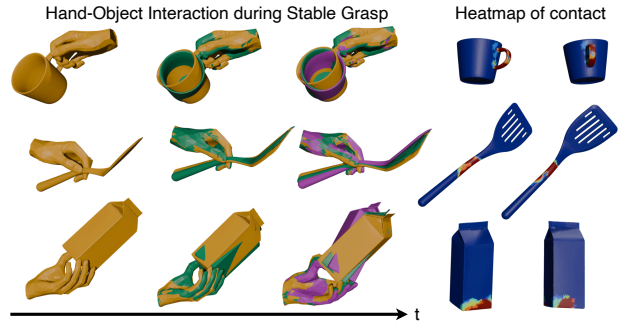

Figure 3. **Stable Grasp Intuition.** Three samples from HOT3D. In each row, we align the hand coordinate system for three frames from one stable grasp. Left: finger articulations and object pose vary over time. Right: contact area (shown as a heat map of objects vertices in contact with the hand) remains consistent.

Similarly if $S_i = \emptyset$, the object is not in contact with the hand and is thus assumed Static. If $S_i \neq \emptyset$ but $\text{IOU}(S_i, S_j) \leq \tau$, the object is in an Unstable Contact.

3.2. ROHIT task and Notations

Assumptions. ROHIT aims to jointly estimate the object poses over hand interaction alone one Hand Interaction Timeline (HIT). This allows reconstructing 3D mesh of the hand and the object, per frame. We assume the knowledge of object category CAD model, as well as the hand-side (left/right) associated with each segment. We also assume the temporal boundaries within the HIT, *i.e.* the start/end of each segment. Note that the assumption of hand side and in-hand segment boundary is implicit in *all prior works* that assume the hand is already grasping the object. These annotations expect the durations (start-end) of the hand grasps as input. Relaxing the assumptions is discussed in the supp.

Following prior works [6, 29, 49], we use MANO [55] to represent the **hand mesh**, which takes as input the per-frame finger articulation vector $\theta^n \in \mathbb{R}^{45}$ and outputs the hand mesh with vertices $V_h^n = \text{MANO}(\theta^n) \in \mathbb{R}^{778\times3}$ in the *hand coordinate system*. When unavailable, we utilise an off-the-shelf method [50] to obtain finger articulations for each frame θ^n. Additionally, we utilise the hand-to-camera (*h2c*) pose T_{h2c}^n, which determines the hand wrist orientation and position. T_{h2c}^n is used to transform meshes from the hand coordinate system to the *camera coordinate system* for each frame.

For the **object mesh**, $V_o \in \mathbb{R}^{|V_o|\times3}$ denotes the known object vertices in the *object coordinate system*. We denote the object-to-hand ($o2h$) poses T_{o2h}^n and the object scale $s \in \mathbb{R}$, which transform the object vertices to $V_{o:h}^n$ in the *hand coordinate system* for each frame. Given the hand-to-camera pose T_{h2c}^n, the object mesh in the camera coordinate is represented as

$$V_{o:c}^n = T_{h2c}^n\,(T_{o2h}^n(s * V_o)) \tag{2}$$

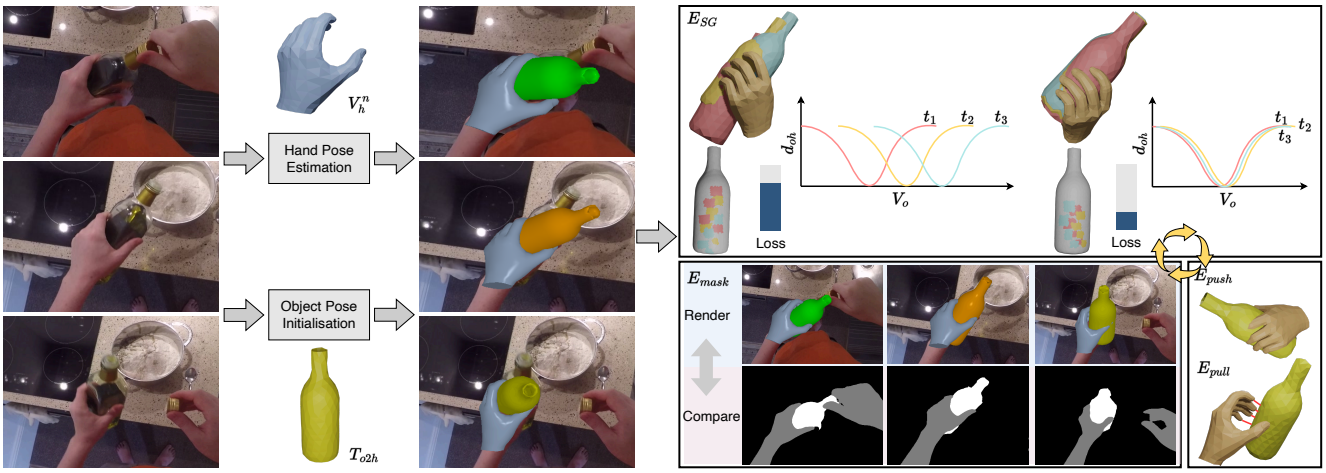

Figure 4. **Optimising the Stable Grasp segment**. We show three frames within one stable grasp. We utilise HaMeR [50] to reconstruct the hand mesh in-the-wild. We initialise T^n_{o2h} to one T_{o2h}, but keep the diverse finger articulations from the hand pose estimates. During optimisation, we measure the distance between each hand vertex and object vertices V_o. In the left plot, we show that the contact area $d_{oh} \approx 0$ differs over time (visualised on the gray bottle). The novel loss, E_{SG}, minimises the variation of distance between the hand and object vertices over time, by adjusting the object's pose relative to the hand. As E_{SG} is minimised, the contact area is aligned (see updated plot). Additional losses are used to regularise the optimisation: E_{mask} renders the reconstruction and compares it against estimated object masks while E_{push} and E_{pull} respectively ensure the object is not penetrated by or away from the hand.

Lastly, to reconstruct the HIT in world coordinates, we use the world-to-camera pose T^n_{w2c}. In Static segments (Sec. 4.1) this allows us to represent the object mesh in the camera coordinate as

$$V^n_{o:c} = T^n_{w2c}(T^n_{o2w}(s * V_o)) \tag{3}$$

where T^n_{o2w} is object-to-world pose. In Stable Grasp and Unstable Contact segments, the hand-to-world pose T^n_{h2w}, obtained via $T^n_{h2w} = (T^n_{w2c})^{-1} \times T^n_{h2c}$, is used to convert between the hand coordinate and the world coordinate,

$$T^n_{o2w} \underset{T^n_{h2w}}{\overset{(T^n_{h2w})^{-1}}{\rightleftarrows}} T^n_{o2h} \tag{4}$$

The world-to-camera pose $T^n_{w2c} = (T^n_{c2w})^{-1}$ is provided from egocentric datasets [3, 62] or can be estimated [57].

4. Constrained Optimisation and Propagation

We propose the Constrained Optimisation and Propagation (COP) framework to reconstruct hand-object mesh pairs along the HIT. Our proposal stems from the understanding of the various constraints governing the changing in-contact vertices of the object along the HIT (Fig 1). To accommodate variable length segments, we always use N sampled frames per segment. Once each segment is optimised, using specific constraints, the pose is propagated to initialise the next segment – as object poses have to remain temporally smooth. We first explain the constrained optimisation for each type of segment.

4.1. Optimising a Static Segment

While the object appears moving in the camera due to head motion in an egocentric video, during the Static segment, the object is stationary in the world coordinate system.

Static Constraint. When the object is static, the object-to-world pose T_{o2w} remains fixed across all N frames in the segment, $T_{o2w} = T^n_{o2w}$. We use Eq. (3) to get the object mesh in camera coordinate $V^n_{o:c}$, and optimise T_{o2w} and the scale s using the *render-and-compare* loss.

Render-and-Compare Loss (E_{mask})**:** This loss focuses on estimating a reconstruction that best matches the 2D projections of object masks throughout the sequence. We measure the error via sum of pixel differences:

$$E^n_{mask} = |\mathcal{C}^n_o \otimes (\mathcal{M}^n_o - \Pi(V^n_{o:c}))|^2_2 \tag{5}$$

where $\mathcal{M}^n_o$ is the object mask which we use for supervision and $\Pi(\cdot)$ is the differentiable projection function [36]. $\mathcal{C}^n_o$ is the occlusion-aware mask as in [29, 79] which only computes the error within regions of the object that are not occluded, set to 1 for the object and the background, and 0 for the hand. This masking avoids penalising the missing parts of the object due to hand occlusion.

4.2. Optimising a Stable Grasp Segment

In a stable grasp, the object's pose is controlled by the in-hand contact vertices. We thus optimise the object relative pose w.r.t. the hand, *i.e.* T^n_{o2h}. This is different from optimising the object w.r.t. to the camera [29, 49]. We use Eq. (2) to get the object mesh in camera coordinate frame $V^n_{o:c}$, and optimise T^n_{o2h} and s.

Stable Grasp Constraint. Following the formal definition from 3.1, we use the stable contact area as our constraint to optimise this segment type. Recall that the object pose relative to the hand can change as long as the vertices in contact with the hand remain stable. We introduce a new loss E_{SG} to model this constraint.

Stable Grasp Loss (E_{SG}): First, we limit this to the hand vertices that are typically in contact with objects – these are the five fingertips (see supp. for visualisation). More formally, let's denote this subset of hand vertices as $V_F \subset V_h$. For each object vertex $v_o \in V_o$, we calculate the distance d_{oh} to each $v_h \in V_F$. We then minimize the average variation of this distance across all pairs of frames n and m:

$$E_{SG} = \sum_{v_o \in V_o} \sum_{v_h \in V_F} \sum_{n=1}^{N} \sum_{m=1}^{N} |d_{oh}^n - d_{oh}^m|_1 \tag{6}$$

$$d_{oh}^n := |v_o^n - v_h^n|_2^2 \tag{7}$$

where v_h refers to one hand vertex, with v_h^n representing its location at frame n; similarly, v_o and v_o^n denote the object vertex and its frame-specific location, respectively.

Note that we assume a rigid object, so we can only minimise E_{SG} by updating its overall pose – i.e. translating and rotating the object relative to the hand. As E_{SG} is minimised, the object's pose is optimised so as to minimise the difference between d_{oh}^n and d_{oh}^m for all pairs of frames. Optimising for this distance is the same as aligning the contact area – *i.e.* if two frames have the same hand-object vertex distance, then the contact area will undoubtedly be aligned.

Figure 4 visualises this loss by considering three timestamps for the hand and bottle (t_1, t_2, t_3). We consider a single hand vertex and plot the distance to all object vertices over time. Before optimising the E_{SG} loss, the distances to bottle vertices changes over time – visualised through the d_{oh} coloured curves (left graph). After optimising for E_{SG}, the plots are better aligned (right graph). We visualise the contact area on the bottle before/after minimising E_{SG}.

In addition to this newly introduced loss, we also use E_{mask}, introduced in Eq 5, as in the Static segment enforces the reconstruction to match the observation of the object in the images. We also use two standard physical heuristic losses E_{push} and E_{pull} to ensure contact and avoid mesh penetration between the hand and the object.

Push (E_{push}) and Pull (E_{pull}) Loss : Motivated by previous works [6, 29, 49, 70], E_{push} pushes the object out of the penetrating region against the hand while the balancing loss E_{pull} pulls the object to touch the hand. For calculations of E_{push} and E_{pull}, refer to the supp.

Combining the four losses, the objective function for the Stable Grasp segment becomes:

$$E(\{T_{o2h}^n\}, s) = \lambda_1 E_{SG} + \sum_{n=1}^{N} (E_{mask}^n + \lambda_2 E_{push}^n + \lambda_2 E_{pull}^n) \tag{8}$$

where λ_1 is the weight for E_{SG} and λ_2 is the weight for E_{push}^n and E_{pull}^n and s is the object scale.

The stable grasp optimisation is overviewed in Fig. 4. E_{SG} is key to optimising a Stable Grasp as it optimises jointly across frames. Note that prior work [29] does add temporal smoothing over time, which we experimentally show to be insufficient for accurate optimisation.

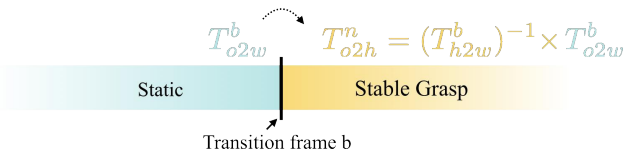

Figure 5. Sample propagation (e.g. from Static to Stable Grasp).

4.3. Optimising an Unstable Contact Segment

When the object is beyond the stable grasp, but still in the hand, we only make assumption on the contact with the hand. We thus utilise E_{mask}, E_{push}, E_{pull} losses *without* the stable grasp assumption, *i.e.* set $\lambda_1 = 0$ in Eq. (8).

4.4. Pose Propagation in COP

When we optimise along the HIT, each segment is optimised in order. Figure 5 illustrates an example propagation from Static to Stable Grasp segment. Once the segment is optimised, we obtain the object-to-world pose at the last frame, b, which is the transitioning pose to the next segment T_{o2w}^b. Intuitively, the object-to-world pose should be consistent at the transition frame b. Where needed, Eq. (4) is utilised to convert from the hand coordinate into the world coordinate, or back from the world to the hand coordinate.

The optimal pose at b is passed to the next consecutive segment as an additional initialisation. Propagating the object-to-world pose at the transition is key to our proposed propagation. Notice that both the pose of the object and its scale are used to initialise the next segment. We propagate between **all consecutive segments** regardless of the type, *e.g.* Stable Grasp → Static or Static → Unstable Contact.

Each initialisation, including the propagated one, is optimised independently based on the constrained losses for the segment type (Sec 4.1-4.3). We select the pose with the minimum E_{mask} as our method's prediction. This is again passed to the next segment in the HIT, and so on.

5. Experiments

5.1. Dataset

With the definition of HIT and Stable Grasp in Sec. 3.1, we annotate timelines from unscripted egocentric videos. In

Table 1. **Dataset Comparison**. Here we compare various characteristics and labels provided by various datasets. We also show statistics of Stable Grasp and HIT (when available). †: subjects in the released train/val set

Dataset	Year	Characteristics			Labels			Stable Grasps' Stats					HIT's Stats			
		In-the-wild	Funct. Intent	Ego	Pose GT	Stable Grasp	HIT	#Env	#Sub	#Cat	#Inst	#Seq	Avg. Duration	#frames	Avg. Seg. Per HIT	#Seq
HOI4D [43]	2022	✗	✓	✓	3D	✗	✗	610	9	20	800	5,000	-	-	-	-
ARCTIC [20]	2023	✗	✓	✓	3D	✗	✗	1	9†	11	11	339	-	-	-	-
HOGraspNet [13]	2024	✗	✗	✓	3D	✓	✗	1	99	30	30	∼3861	-	-	-	-
HOT3D [3]	2024	✗	✓	✓	3D	✗	✗	4	19	33	33	295	-	-	-	-
HOT3D-HIT (ours)	2025	✗	✓	✓	3D	✓	✓	4	9	22	22	1,239	121.1s	410,650	29.1	113
MOW [6, 49]	2021	✓	✓	✗	✗	✗	✗	500	500	121	500	500	-	-	-	-
EPIC-HIT (ours)	2025	✓	✓	✓	2D Mask	✓	✓	141	31	9	∼390	2,431	13.8s	79,736	2.8	96

HOT3D [3], we automatically annotate HITs from the 3D ground truth – we refer to these sequences as **HOT3D-HIT**. Object masks are provided by the ground truth. In EPIC-KITCHENS [16], we manually annotate HIT segments for 9 rigid and commonly used object categories (plate, bowl, bottle, cup, mug, can, pan, saucepan, glass). Corresponding object masks are available from the VISOR dataset [18]. We refer to this datasets as **EPIC-HIT**.

We compare against existing datasets in hand-object reconstruction in Tab. 1 (full table in supp.). For HOT3D, we automatically extract 1,239 stable grasps and extend them to 113 HITs covering 3,288 segments (w/ 872 Static and 1177 Unstable Contact). For EPIC-KITCHENS, we labelled 2,431 video clips of Stable Grasp from 141 distinct videos in 31 kitchens. We additionally manually label 96 HITs, covering 296 segments (135 Static, 106 Stable Grasp, and 28 Unstable Contact). Details in supp.

5.2. Implementation Details

Following HOMan [29], we sparsely and linearly sample 30 frames from each segment for optimisation and evaluation. For each object, we use 10 rotation initialisations and 1 global translation to initialise T_{o2h}, following [29]. The error E_{mask} is defined in pixels whilst E_{SG}, E_{push} and E_{pull} are defined in 3D space (metres). We use the camera's focal length f as a scaling factor: $\lambda_f = f * \texttt{render_size}$. We use $\texttt{render_size} = 256$ and set $\lambda_1 = \lambda_f$ and $\lambda_2 = 0.1 * \lambda_f$ (Eq 8). The optimisation takes on average 30s on a RTX 4090 for one 30-frame segment. Please refer to the supp. for additional implementation details.

5.3. Baselines and Quantitative Metrics

We focus on baselines that are able to reconstruct object poses in the hand, using a predefined CAD model. Since COP does not require training, we do not compare with data-driven methods [20, 41, 41, 70, 71]. We compare to:

- **HOMan** [29]—A common CAD-based baseline that progressively optimises the object pose relative to the hand. HOMan implements temporal smoothing and uses E_{pull} and E_{push}. Similar to our method, we use VISOR [18] masks for fair comparison.
- **Rigid** [60]—assumes that objects are not allowed any motion within grasp, minimising overall rotation and translation of the object relative to the hand.
- **Dynamic**—A variation of COP without the stable grasp constraint ($\lambda_1 = 0$).

We note that RHOV [49] is a relevant baseline but no code is available to be used as a baseline. CAD-agnostic methods [21, 53, 68, 73, 74] fail catastrophically on our datasets (see supp) making them unsuitable as baselines.

We use two standard metrics for evaluation with/without 3D ground truth, suitable for all segment types:

Average Distance (ADD) is the standard metric for methods with 3D ground truth. Following [13, 30, 37, 69], we measure the distance of corresponding vertices between GT and predicted object vertices, and average it over vertices and frames. ADD is 1 for a sequence if average distance is less than 10% of the object's diameter, and 0 otherwise. In symmetric CAD models, we calculate the minimal average distance among the valid symmetric transformations [31].

Intersection-over-Union (IOU). We use IOU as a proxy of pose accuracy when 3D GT is *not* available. We measure IOU between ground truth mask and rendered mask for the object in camera view. We report average IOU across all frames. In case of occlusion with other components, only the non-occluded area of the rendered projection is used.

We also propose variations of these standard metrics that particularly measure the Stable Grasp in HIT:

Average Stable Contact Area at ADD Success (SCA-ADD). When a pose is considered correct for a sequence, *i.e.* average distance is within 10% of the object's diameter and thus ADD is 1, we measure the stable contact area across the sequence, defined as the average IOU of in contact area between all pairs of frames (Sec. 3.1). SCA-ADD is set to 0 when ADD is 0 (average distance below threshold). We average SCA-ADD over all examples. We use this metric to showcase our ability to reconstruct stable grasps.

Average Stable Contact Area at high IOU (SCA-IOU). We analogously report SCA when IOU is more than certain thresholds. We use 80% as threshold on HOT3D, and report both 80% and 60% on EPIC. SCA-IOU is set to 0 when IOU is below the threshold. We report average SCA-IOU.

IOU vs. SCA-IOU. Note that IOU and SCA-IOU can be contradictory. A method can maximise IOU by individually fitting to each mask, resulting in a lower SCA-IOU. Using

Table 2. **Results on Stable Grasp in HOT3D-HIT.** Green for best and yellow for second best. COP† is COP w/o propagation.

Category	SCA-IOU				ADD				SCA-ADD			
	HOMan [29]	Rigid [60]	Dynamic	COP†	HOMan [29]	Rigid [29]	Dynamic	COP†	HOMan [29]	Rigid [60]	Dynamic	COP†
bottle_bbq	23.1	39.2	34.2	43.5	4.9	30.5	52.4	57.3	3.8	16.9	21.0	26.7
bottle_mustard	15.2	16.8	23.2	32.3	0.0	12.5	37.5	37.5	0.0	5.9	14.9	18.3
bottle_ranch	16.8	33.3	32.9	40.4	9.1	31.8	50.0	56.8	5.9	17.4	19.6	25.4
bowl	39.4	53.4	32.8	48.7	36.5	84.9	84.9	89.7	23.3	44.8	28.5	43.7
can_parmesan	29.1	44.8	32.8	43.5	0.0	19.0	27.0	30.2	0.0	9.0	10.4	14.7
can_soup	32.4	50.3	37.1	50.6	3.0	14.1	21.2	26.3	2.5	6.7	7.6	13.8
can_tomato_sauce	36.2	46.3	33.2	45.8	2.0	2.9	9.8	9.8	1.6	1.1	3.2	4.2
carton_milk	24.9	27.5	29.8	38.3	0.0	35.3	17.6	29.4	0.0	15.8	6.0	14.5
carton_oj	5.5	35.0	25.6	39.3	0.0	31.2	43.8	43.8	0.0	16.2	14.4	20.6
cellphone	49.3	8.3	17.6	23.0	42.3	9.6	23.1	25.0	31.6	6.7	8.5	14.0
coffee_pot	27.1	42.4	35.5	43.7	18.4	57.1	71.4	75.5	15.3	29.1	28.1	35.2
dino_toy	24.8	15.9	38.6	40.5	33.3	66.7	72.2	77.8	24.8	32.7	34.5	39.7
food_vegetables	43.0	41.6	40.7	48.9	9.5	23.8	33.3	23.8	8.4	14.1	18.7	13.2
keyboard	37.6	31.8	21.9	31.4	60.9	84.1	92.8	91.3	35.6	29.4	19.8	28.1
mouse	48.3	59.3	44.5	56.3	18.6	39.5	67.4	72.1	13.8	28.9	34.1	43.8
mug_white	15.2	33.1	30.9	42.4	11.1	28.9	37.8	57.8	8.1	15.9	17.5	28.9
plate_bamboo	36.6	56.3	33.6	55.1	35.6	81.4	89.8	93.2	24.4	50.0	30.5	50.5
potato_masher	2.2	5.3	17.7	26.5	4.9	50.5	47.6	29.4	2.9	28.3	15.4	25.2
puzzle_toy	37.1	46.4	32.1	44.8	24.4	54.9	51.2	64.6	17.9	29.2	19.1	31.7
spatula_red	1.3	14.6	26.7	35.4	4.8	50.8	76.2	85.7	3.3	33.0	24.6	35.4
whiteboard_eraser	23.0	42.6	36.0	42.1	20.0	80.0	100.0	100.0	9.9	37.8	36.0	42.1
whiteboard_marker	30.4	49.2	37.4		0.0	100.0	66.7	45.7	0.0	71.2	27.3	35.4
Average	27.8	37.4	30.8	**42.0**	16.8	43.0	51.9	**58.1**	11.5	22.9	18.4	**26.8**

Table 3. **Results on Stable Grasp in EPIC-HIT.** Green for best and yellow shows the second best. COP† is COP w/o propagation.

Category	IOU				SCA@0.8				SCA@0.6			
	HOMan [29]	Rigid [60]	Dynamic	COP†	HOMan [29]	Rigid [60]	Dynamic	COP†	HOMan [29]	Rigid [60]	Dynamic	COP†
bottle	56.6	66.0	75.5	72.2	3.5	16.4	21.8	29.6	5.7	51.9	36.6	56.3
bowl	54.2	52.9	57.9	55.1	1.0	10.5	9.3	11.9	1.6	28.2	19.7	32.0
can	47.5	49.8	55.2	53.6	3.5	12.7	13.0	16.5	5.3	24.1	19.4	28.1
cup	56.9	61.4	67.5	66.1	5.5	12.4	16.5	20.0	7.8	43.2	38.9	49.7
glass	55.4	57.6	65.2	63.1	3.0	10.0	14.8	15.9	4.2	36.3	30.1	40.9
mug	59.4	58.1	63.8	62.1	3.9	6.5	6.3	9.2	6.6	36.3	29.6	42.5
pan	48.3	45.8	48.9	42.4	0.5	4.5	4.3	3.5	1.9	18.6	12.8	15.1
plate	61.1	61.5	68.1	65.3	1.0	15.5	17.6	22.5	1.6	39.3	30.1	46.4
saucepan	51.1	53.1	57.5	54.4	0.4	3.1	5.4	5.3	2.4	32.7	25.7	31.1
Average	54.8	55.8	**61.7**	58.2	1.9	10.6	12.2	**15.0**	3.3	33.4	25.2	**36.5**

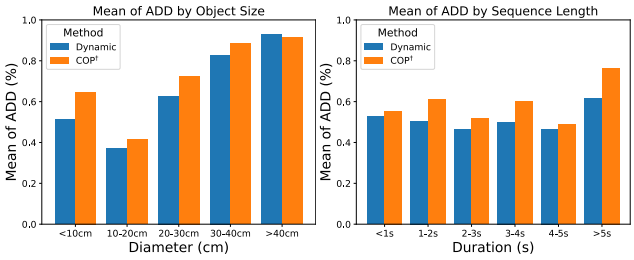

Figure 6. Improvement of COP over Dynamic method for different object sizes (left) and Stable Grasp lengths (right).

both metrics allows us to understand the performance of different baselines versus our proposed COP.

5.4. Results and Ablation

Stable Grasp in HOT3D-HIT. We first study the reconstruction of Stable Grasp segments. To ensure fair comparison, here COP does not employ propagation. Table 2 contains per-category results for Stable Grasp in HOT3D-HIT. On average, COP improves ADD from 51.9% with the dynamic assumption to 58.1% within the Stable Grasp, and improves SCA-ADD from 22.9% with the rigid assumption to 26.8%. Categories like "potato_masher" significantly improve in ADD score (+16.5%). The baseline 'Rigid' has high SCA-ADD, but ADD is significantly lower than COP. HOMan [29] performs badly on most categories as the contact is not maintained through its iterative optimisation.

In Fig. 6, we investigate the impact of Stable Grasp sequence length and different object sizes. COP consistently outperforms Dynamic, only dropping slightly for large objects in HOT3D. Moreover, COP performs significantly better for smaller objects and on long sequences.

Stable Grasp in EPIC-HIT. Table 3 contains analogous per-category results for Stable Grasp in EPIC-HIT. COP outperforms baselines on the SCA metric. Dynamic performs well on IOU, as it fits to each mask individually but *does not* maintain a stable grasp indicated by the lower SCA metric. Note that IOU is the 2D proxy metric of pose accuracy; E_{SG} regularises the IOU fitting to improve SCA. Rigid achieves best SCA results on the "pan" category only – indicating the functional hold of the pan does not allow finger articulations. Figure 7 shows reconstruction results of COP on Stable Grasp in EPIC-HIT.

HIT in HOT3D-HIT. In Tab. 4, we show quantitative results on the full timelines (HIT) reconstruction on HOT3D-HIT. While propagation could theoretically improve the Dynamic baseline, we focus our propagation ablation on COP. Improvements can be seen across the board – ADD in Stable Grasp segments is improved by 12.2% and Static segments

Figure 7. **Qualitative Results on Stable Grasp in EPIC-HIT** (2 examples/category): projected reconstruction results and reconstruction in rotated views. Bottom: failure cases due to wrong hand pose (left) and extreme occlusion (right).

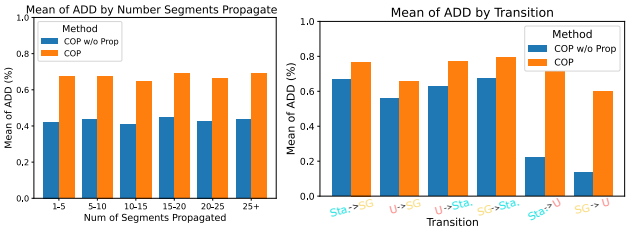

Figure 8. Impact of propagation on segment count and types.

benefits also, improving by 14.5%. This highlights that COP with propagation improves the reconstruction across all segment types.

Figure 8 summarises the impact of propagation on HITs with different segment count and types. Improvement is consistent across counts. Notably, transition to Unstable Contact shows significant improvement in ADD.

HIT in EPIC-HIT. In Tab. 5 we report analogous results on EPIC-HIT. COP again shows consistent improvements over all metrics across different types of segments. We show qualitative results for HIT reconstruction in Fig. 2.

Ablation on the constraint in Eq. (6). Recall that in E_{SG}, we first index every object vertex $v_o \in V_o$, compute its distance to each finger tips vertex, then constrain the variation over all pairs of frames. In Tab. 6, we ablate two variants, aiming to maximise the stable contact area: i) select the nearest object vertex v_o^* corresponds to each $v_h \in V_F$, instead of selecting all object vertices V_o; ii) select N consecutive frames to constrain d_{oh}. The results show that our constraint performs the best.

Table 4. Results on HOT3D-HIT

Method	Stable Grasp		Static	Unstable Contact
	ADD	SCA-ADD	ADD	ADD
COP *w/o Propagation*	58.1	26.8	64.7	17.3
COP	**70.3**	**34.1**	**79.2**	**67.1**

Table 5. Results on EPIC-HIT

Method	Stable Grasp		Static	Unstable Contact
	IOU	SCA@0.8	IOU	IOU
COP *w/o Propagation*	68.0	23.8	65.8	74.3
COP	**69.2**	**25.5**	**68.3**	**79.0**

Table 6. Ablation on the Stable Grasp Loss E_{SG} variants on HOT3D. We show improvement over the Dynamic Baseline

Obj. Vert. Selection	Frames Selection	ADD	SCA-ADD
V_o	N^2 pairs	58.1 (+6.2)	26.8
v_o^*	N^2 pairs	52.9 (+1.0)	21.3
V_o	N consecutive	54.2 (+2.3)	22.5
Dynamic Baseline		51.9 (+0.0)	18.4

Table 7. **Ablation on the loss weights.** Chosen λ_1 and λ_2 in blue

λ_1	λ_2	ADD	SCA-ADD
0	0.1	51.9	18.4
1	0	54.8	25.0
1	0.1	58.1	26.8
10	0.1	51.5	26.6

Ablation on the weights in Eq. (8). In Tab. 7, we ablate the weights λ_1 and λ_2 introduced in the loss function Eq. (8). The results suggest that E_{SG} is important while the physical heuristics E_{pull} and E_{push} are also necessary. Dynamic is equivalent to $\lambda_1 = 0$ and $\lambda_2 = 0.1$.

Additional Results. In supp., we (i) ablate the robustness to noise in the segment boundaries along the HIT; (ii) report results on stable grasp segments from the ARCTIC dataset [20] with matching conclusions; and (iii) show further examples of failure of CAD-Agnostic methods on in-the-wild recordings.

6. Conclusion

We proposed the Reconstructing Objects along Hand Interaction Timelines (ROHIT) task – which aims to reconstruct an object along time including when the object is static in the scene, in an stable or unstable grasp. To tackle the task, we propose the Constrained Optimisation and Propagation (COP) framework which builds around Stable Grasp and propagates poses across segments for superior reconstruction. We propose HOT3D-HIT (with 3D ground truth) and EPIC-HIT (in-the-wild) datasets to evaluate COP. We highlight the efficacy of stable grasp and pose propagation on both datasets. By reconstructing full timelines, we hope to encourage future works to quantitatively evaluate timeline reconstruction methods in-the-wild.

Acknowledgements This work is supported by EPSRC UMPIRE EP/T004991/1. Z Zhu is supported by UoB-CSC scholarship. S Bansal is supported by a Charitable Donation to the University of Bristol from Meta. S. Tripathi is supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). We thank Ahmad Darkhalil for help with VISOR masks.

References

[1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1001–1010, 2023. 2

[2] Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. In *Thirteenth International Conference on 3D Vision*, 2025. 2

[3] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7061–7071, 2025. 1, 4, 6, 13, 14, 15

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Proceedings of the European Conference on Computer Vision*, pages 361–378, 2020. 1, 13, 15

[5] Ian M. Bullock, Raymond R. Ma, and Aaron M. Dollar. A hand-centric classification of human and robot dexterous manipulation. *IEEE Transactions on Haptics*, 6(2):129–144, 2013. 3

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12417–12426, 2021. 1, 2, 3, 5, 6, 15

[7] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 13, 15

[8] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. MonoRUn: Monocular 3D object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 2

[9] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 17

[10] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *Proceedings of the European Conference on Computer Vision*, pages 231–248, 2022. 1, 2

[11] Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: Monocular hand-held objects reconstruction with transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2025. 2

[12] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4769–4778, 2023. 2

[13] Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, and Tae-Kyun Kim. Dense hand-object(ho) graspnet with full grasping taxonomy and dynamics. In *Proceedings of the European Conference on Computer Vision*, 2024. 6, 15

[14] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 14

[15] M.R. Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on Robotics and Automation*, 5(3):269–279, 1989. 3

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*, pages 720–736, 2018. 6, 14

[17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 13, 14

[18] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems*, 2022. 6, 14

[19] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In *Advances in Neural Information Processing Systems*, pages 2127–2160, 2024. 2

[20] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6, 8, 13, 14, 15

[21] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 2, 6, 13, 16, 17

[22] Thomas Feix, Javier Romero, Heinz Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The GRASP Taxonomy of Human Grasp Types. *IEEE Transactions on Human-Machine Systems*, 46(1):66–77, 2016. 3

[23] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 15

[24] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9784–9794, 2019. 2

[25] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 516–532, 2022. 2

[26] 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 Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2020. 13, 15

[27] 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 Conference on Computer Vision and Pattern Recognition*, 2023. 2

[28] Yana Hasson, Gül 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 Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 2, 15

[29] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *International Conference on 3D Vision (3DV)*, pages 659–668, 2021. 1, 2, 3, 4, 5, 6, 7, 14, 16

[30] Tomas Hodan, Daniel Barath, and Jiri Matas. EPOS: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11703–11712, 2020. 6

[31] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. BOP challenge 2020 on 6D object localization. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 577–594. Springer, 2020. 6

[32] 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 *Proceedings of SIGGRAPH Asia 2022 Conference Papers*, 2022. 2, 13

[33] Elkhan Ismayilzada, MD Khalequzzaman Chowdhury Sayem, Yihalem Yimolal Tiruneh, Mubarrat Tajoar Chowdhury, Muhammadjon Boboev, and Seungryul Baek. Qortformer: Query-optimized real-time transformer for understanding two hands manipulating objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3895–3903, 2025. 2

[34] 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

[35] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 1, 2

[36] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018. 4

[37] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 954–962, 2015. 6

[38] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10138–10148, 2021. 2, 13, 15

[39] Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. HHMR: Holistic Hand Mesh Recovery by Enhancing the Multimodal Controllability of Graph Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 645–654, 2024. 2

[40] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[41] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12989–12998, 2023. 2, 6

[42] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 2

[43] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 6, 15

[44] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7037–7047, 2025. 2

[45] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. AutoShape: Real-time shape-aware

monocular 3D object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15621–15630, 2021. 2

[46] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26, 2017. 15

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 17

[48] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. Foundpose: Unseen object pose estimation with foundation features. In *Proceedings of the European Conference on Computer Vision*, 2024. 2, 17

[49] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022. 1, 2, 3, 4, 5, 6, 15

[50] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 2, 3, 4, 14

[51] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *2025 International Conference on 3D Vision (3DV)*, 2025. 14

[52] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pages 12242–12254, 2025. 2

[53] Aditya Prakash, Matthew Chang, Matthew Jin, Ruisen Tu, and Saurabh Gupta. 3d reconstruction of objects in hands without real world 3d supervision. In *Proceedings of the European Conference on Computer Vision*, 2024. 2, 6

[54] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[55] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Trans. Graph*, 36:17, 2017. 3

[56] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1749–1759, 2021. 2

[57] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113. IEEE Computer Society, 2016. 4

[58] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 15

[59] Edgar Sucar, Kentaro Wada, and Andrew Davison. NodeSLAM: Neural Object Descriptors for Multi-View Shape Reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 13

[60] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. Showme: Benchmarking object-agnostic hand-object 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1935–1944, 2023. 6, 7, 13, 15

[61] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision*, pages 581–600, 2020. 1, 13, 15

[62] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Advances in Neural Information Processing Systems*, 2023. 4, 14

[63] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1674, 2022. 2

[64] He Wang, Srinath Sridhar, Jingwei Huang, Julien P. C. Valentin, Shuran Song, and L. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[65] Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction, 2024. 15

[66] Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5735–5745, 2023. 2

[67] Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. Magichoi: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5957–5968, 2025. 2

[68] Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d. *arXiv preprint arXiv:2404.06507*, 2024. 6

[69] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for

6D object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems*, 2018. 2, 6

[70] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11097–11106, 2021. 1, 2, 5, 6

[71] 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 Conference on Computer Vision and Pattern Recognition*, pages 2750–2760, 2022. 2, 6

[72] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 15

[73] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022. 1, 2, 6

[74] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19717–19728, 2023. 2, 6, 13, 16, 17

[75] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 16, 17

[76] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7072–7084, 2025. 2

[77] Zhenjun Yu, Wenqiang Xu, Pengfei Xie, Yutong Li, Brian W Anthony, Zhuorui Zhang, and Cewu Lu. Dynamic reconstruction of hand-object interaction with distributed force-aware contact representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8590–8599, 2025. 2

[78] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural Manipulation Synthesis with a Hand-Object Spatial Representation. *ACM Transactions on Graphics*, 40(4), 2021. 1, 13

[79] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 34–51, 2020. 4

[80] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2024. 2

Reconstructing Objects along Hand Interaction Timelines in Egocentric Video

Supplementary Material

A. Overview

On the project's webpage, we provide qualitative videos showcasing qualitative results and include details describing the video in Sec. B. Rest of this document is arranged as follows. Section C provides additional annotation details of the EPIC-HIT and HOT3D-HIT datasets. We ablate the robustness of COP to the boundaries of segments in HIT in Sec D. Results on Stable Grasp in the ARCTIC dataset [20] are provided in Sec. F. Additional implementation details are provided in Sec. G. We then qualitatively evaluate CAD-agnostic models in Sec. H. Finally, in Sec. I, we discuss limitations of our work.

B. Qualitative Video

We include videos showcasing the reconstruction results on the two datasets using our proposed approach COP. The video collection contains examples from both EPIC-HIT and HOT3D-HIT. In each case, we show the original video (left), projected reconstruction in camera frame (middle) and 3D hand-object reconstruction from 2 different views (right). We also show the object and hands in world coordinate frame (bottom) with camera pose as a red prism.

Additionally, we provide examples of Stable Grasp sequences in EPIC-HIT. There are two examples from each object category (bottle, can, mug, glass, bowl, cup, plate, pan, saucepan).

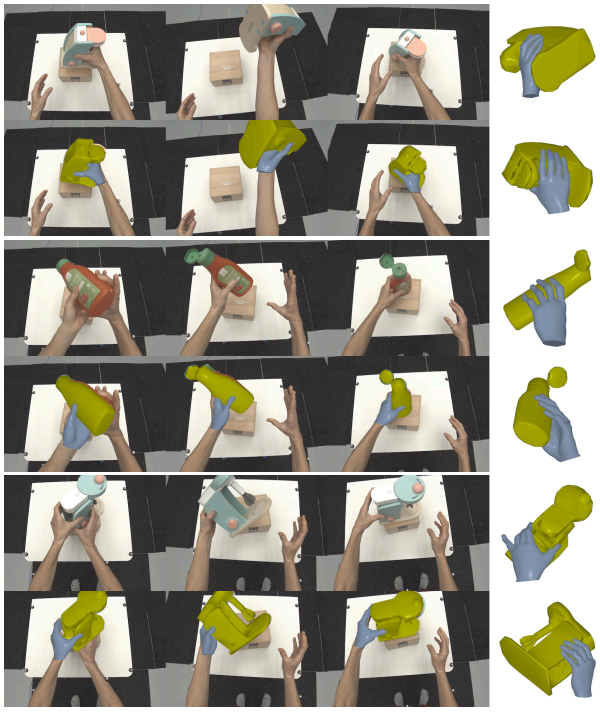

Figure 9. Qualitative results of COP on Stable Grasp from ARCTIC [20]. There are three sequences visualised here. Top row in each sequence contains input frames. Bottom row in each sequence contains frames with reconstructed hand and object. Last column shows the hand and object reconstruction from two different perspectives.

C. Annotating HOT3D-HIT and EPIC-HIT

With the definition of HIT and Stable Grasp in Section 3.1 of the main paper, we annotate Hand-Interaction Timelines in two datasets.

C.1. HOT3D-HIT.

For the HOT3D [3] dataset which has 3D ground truth, we automatically extract stable grasps sequences with threshold $\tau = 0.5$ in Equation 1 in the main paper. We locate $1,239$ stable grasps sequences which we then extend automatically to HIT using the annotations to identify when the object is in-view. In total, we label 113 HITs covering $410,650$ frames across 20 videos, $3,288$ segments (872 Static, 1239 Stable Grasp, 1177 Unstable Contact) and 22 objects.

C.2. EPIC-HIT.

We annotate the temporal segments of HIT from the EPIC-KITCHENS [17] videos. This offers a dataset distinct from prior works, which are collected in lab settings [4, 61, 78] or contain recordings specifically collected to evaluate grasps with no underlying action [7, 26, 38]. Instead, we aim to leverage Stable Grasp definition to identify HIT sequences within unscripted egocentric videos of daily actions. Note that we exclude interactions with non-rigid objects and only focus on interactions with rigid known objects. We next detail our annotation pipeline:

1. Identifying candidate clips. The ultimate goal of hand-object reconstruction is to generalize to any rigid or dynamic objects, including those belonging to novel classes. However, as we show later in Sec. H, current approaches for reconstruction of unknown objects [21, 32, 59, 60, 74] are still in their infancy. We thus restrict our scope to known object categories and focus instead of high-fidelity hand-object reconstruction. Note that this is distinct from assuming instance-level CAD models – the general CAD model of a bottle might not exactly match all bottles in daily life. We exclude tiny objects and shortlist 9 categories frequently

Table 8. Sensitivity to noisy boundary on HOT3D stable grasp subset.

Method	ADD	SCA-ADD
HOMAN [29]	15.0	10.0
COP	70.0	32.9
COP w/ Noisy Stable Grasp Boundaries	60.0	27.4

Figure 10. Eight contact regions: five fingertips V_F + three palm areas. The contact regions serve two purposes: bounding the object inside and attracting the object closer to these regions.

used in kitchens: plate, bowl, bottle, cup, mug, can, pan, saucepan, glass[1]. We use annotations and narrations to find clips where a hand is in contact with one of these categories.

2. Annotating Stable Grasp. Two annotators were asked to label the start-and-end frames following the Stable Grasp definition. We discard segments when, (i) both the hand and object are out-of-view during the sequence or, (ii) the object does not match the category CAD model specified.

In total, we label $2,431$ video clips of stable gasps from 141 distinct videos in 31 kitchens [17]. For each clip, we provide a start and end time of the stable grasp, as well as $319,661$ segmentation masks for the hand and the object during the stable grasp from the dense VISOR annotations [18]. Of these, $1,446$ contain left hand stable grasps and 985 contain right hand stable grasps.

3. Annotating HIT segments. Once we have the stable grasps annotated, we extend them to HIT. We select 42 videos that have verified camera pose estimates from [62] with metric scale and gravity available from [51]. Manual annotations for temporal segments are then added to form consecutive segments labelled with segment type. In total, we label 96 HITs, covering $79,736$ frames and 269 segments (135 Static, 106 Stable Grasp, 28 Unstable Contact).

C.3. Dataset Comparison

Table 9 provides a more comprehensive comparison of our datasets with regularly used datasets for hand-object reconstruction. This is an extension of Table 1 in the main paper.

D. Dependency on Accurate Boundaries

COP relies on the provided HIT segment boundaries. One limitation of the method is the need for accurate start-end times of all segments in the hit. These annotations can be relieved if segments are estimated through a localisation model or VLM given a labelled training dataset. While our results in the paper use labelled segments of Hand Interaction Timeline (HIT), we provide an ablation on the need for accurate segment boundaries.

To assess the sensitivity of COP to labelling boundary accuracy, we add random noise—sampled from $\text{Uniform}(10, 30)$ frames—to the ground-truth boundaries for 40 randomly selected Stable Grasp samples from HOT3D. As shown in Table 8, noisy boundaries leads to a performance drop for COP; however, even with such noise, COP still outperforms the baseline [29] by a large margin.

[1]For **object mesh**, we made per-category CAD model in Blender [14].

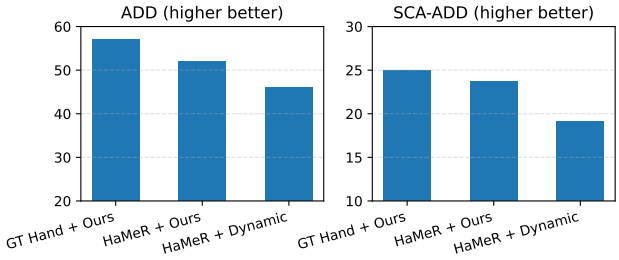

Figure 11. Robustness to noisy hand poses.

E. Sensitivity to Hand Pose Hoise

We analyse our method's sensitivity to the hand pose noise, using a random subset of 100 stable grasp segments from HOT3D. We run HaMeR [50] to obtain the finger poses and use these as input to our method.

Figure 11 compares the results. When we switch from ground-truth to estimated hand poses, our method drops reasonably for both ADD and SCA-ADD metrics. However, our method still clearly outperforms the *best* performing baseline – Dynamic. This finding is in line with the results in the paper.

F. Results on Stable Grasp in ARCTIC

In addition to HOT3D [3] and EPIC-KITCHENS [16], we also explore the ARCTIC dataset [20] with 3D ground truth for HIT reconstruction. However, due to short clips in the dataset, we only evaluate the stable grasp segments on this dataset. Similar to HOT3D, we automatically extract stable grasp sequences with threshold of $\tau = 0.5$ and identify 1303 stable grasp sequences across 9 subjects covering 11 categories. Table 10 contains per-category results for stable grasps in ARCTIC. COP outperforms the baseline [29] and alternate assumptions on all the 11 CAD-model categories. Categories like "capsule machine" see significant improvement in ADD score (+12.6). On average, COP improves ADD from 56.0 with dynamic assumption to 65.1 using the stable grasp assumption.

Table 9. **Dataset Comparison**. Here we compare various characteristics and labels provided by various datasets. We also show statistics of Stable Grasp and HIT (when available). *: object poses or segments are not provided. †: subjects in the released train/val set

Dataset	Year	Characteristics			Labels			Stable Grasps' Stats					HIT's Stats			
		In-the-wild	Funct. Intent	Ego	Pose GT	Stable Grasp	HIT	#Env	#Sub	#Cat	#Inst	#Seq	Avg. Duration	#frames	Avg. Seq. Per HIT	#Seq
FPHA [23]	2018	✗	✓	✓	3D	✗	✗	3	6	4	4	1,175	-	-	-	-
HO3D [26]	2020	✗	✗	✗	3D	✓(part)	✗	1	10	10	10	65	-	-	-	-
ContactPose [4]	2020	✗	✓	✗	3D	✓	✗	1	50	25	25	2,306	-	-	-	-
GRAB [61]	2020	✗	✓	✗	3D	✗	✗	1	10	51	51	1,334	-	-	-	-
H2O [38]	2021	✗	✓	✓	3D	✗	✗	3	4	8	8	24	-	-	-	-
DexYCB [7]	2021	✗	✗	✗	3D	✗	✗	1	10	20	20	1,000	-	-	-	-
HOI4D [43]	2022	✗	✓	✓	3D	✗	✗	610	9	20	800	5,000	-	-	-	-
Assembly101 [58]	2022	✗	✓	✗	3D Hand*	✗	✗	1	53	15	15	4,321	-	-	-	-
OakInk [72]	2022	✗	✓	✗	3D	✗	✗	1	12	32	100	1,356	-	-	-	-
SHOWMe [60]	2023	✗	✗	✗	3D	✓	✗	1	15	42	42	96	-	-	-	-
ARCTIC [20]	2023	✗	✓	✓	3D	✗	✗	1	9†	11	11	339	-	-	-	-
ARCTIC w/ Stable Grasp	2025	✗	✓	✓	3D	✓	✗	1	9	11	11	1,303	-	-	-	-
HOGraspNet [13]	2024	✗	✗	✓	3D	✓	✗	1	99	30	30	~3861	-	-	-	-
HO-Cap [65]	2024	✗	✗	✓	3D	✗	✗	1	9	64	64	64	-	-	-	-
HOT3D [3]	2024	✗	✓	✓	3D	✗	✗	4	19	33	33	295	-	-	-	-
HOT3D-HIT (ours)	2025	✗	✓	✓	3D	✓	✓	4	9	22	22	1,239	121.1s	410,650	29.1	113
Core50 [46]	2017	✓	✗	✗	2D Mask	✗	✗	11	-	10	50	550	-	-	-	-
MOW [6, 49]	2021	✓	✓	✗	✗	✗	✗	500	500	121	500	500	-	-	-	-
EPIC-HIT (ours)	2025	✓	✓	✓	2D Mask	✓	✓	141	31	9	~390	2,431	13.8s	79,736	2.8	96

Table 10. **Results on ARCTIC.** Green shows the best performing method per metric and yellow shows the second best. COP† is COP without propagation.

Category	SCA-IOU				ADD				SCA-ADD			
	HOMan	Rigid [60]	Dynamic	COP†	HOMan	Rigid [60]	Dynamic	COP†	HOMan	Rigid [60]	Dynamic	COP†
box	30.9	67.3	47.5	71.8	33.3	37.7	52.9	60.1	16.9	28.7	26.7	45.5
capsulemachine	34.4	34.9	43.8	66.6	42.1	48.4	45.3	57.9	24.5	34.1	26.1	43.7
espressomachine	36.9	49.2	52.2	73.0	44.6	48.5	64.4	72.3	28.8	36.2	38.2	55.6
ketchup	18.0	32.7	48.4	62.3	15.1	51.9	39.6	56.6	9.3	37.8	23.7	39.2
laptop	35.3	62.0	51.8	69.1	43.8	45.1	60.4	63.9	28.0	37.1	34.8	46.9
microwave	36.1	51.5	48.8	76.1	56.2	50.9	77.7	83.9	27.3	35.3	41.8	64.2
mixer	34.3	37.9	51.1	69.4	45.1	48.4	66.4	73.8	26.7	32.2	38.8	54.2
notebook	38.7	57.8	55.4	66.6	33.8	43.0	53.6	61.6	20.8	33.2	32.4	42.7
phone	39.5	36.7	52.4	62.2	28.1	39.7	34.9	46.6	19.9	29.2	21.4	30.4
scissors	5.2	0.0	16.2	23.7	7.0	47.4	43.9	70.2	5.2	36.3	25.6	47.7
waffleiron	36.3	48.3	51.0	65.0	45.0	59.5	72.5	76.3	26.2	39.2	40.1	51.5
Average	33.1	46.6	49.0	66.1	37.1	46.9	56.0	65.1	22.0	34.2	32.0	46.9

Table 11. Ablation on the Stable Grasp Loss E_{SG} variants on ARCTIC. We show improvement over the Dynamic Baseline

Obj. Vert. Selection	Frames Selection	ADD	SCA-ADD
V_o	N^2 pairs	65.1 (+9.1)	46.9
v_o^*	N^2 pairs	58.4 (+2.4)	37.2
V_o	N consecutive	59.1 (+3.1)	38.1
Dynamic Baseline		56.0 (+0.0)	32.0

Table 12. **Ablation on the weights.** We highlight our choice of λ_1 and λ_2 (blue) on ARCTIC

λ_1	λ_2	ADD	SCA-ADD
0	0.1	56.0	32.0
1	0	63.0	45.0
1	0.1	65.0	47.0
10	0.1	49.0	39.0

Figure 9 shows qualitative results on Stable Grasp from ARCTIC. In Tab. 11 similar to the ablation in the main paper for HOT3D-HIT, we ablate Stable Grasp Loss E_{SG} and show improvement over the Dynamic baseline. Furthermore, in Tab. 12, we ablate the weights on ARCTIC and draw similar conclusion as the analogous ablation on the HOT3D dataset (Table 7 in the main paper).

G. Additional Implementation Details

Physical Loss E_{push} and E_{pull}. In the main paper, we note our usage of physical repulsion and attraction losses E_{push} and E_{pull}. These are similar to the repulsion and attraction losses in [28].

The term E_{push} ensures all object vertices are located in-

side the contact surface of the hand (Fig. 10). E_{push} applies independently to each frame, hence we omit the superscript ^{n}. For each $v_o \in V_o$, we locate the nearest vertex in hand contact regions, and compute the distance along the surface normal of this hand vertex. Object vertices that penetrate into the contact surface will have negative values. We maximise those negative values, truncating the positive ones:

$$E_{push} = \sum_{v_o \in V_o} -1 * \min(d_v, 0) \tag{9}$$

$$d_v := \langle v_o - v_h^*, n_h^* \rangle \tag{10}$$

where v_h^* is the corresponding nearest vertex on the hand and n_h^* is the surface normal of v_h^*.

In addition to E_{push}, which pushes the object out of the penetrating region against the hand, we use a balancing loss E_{pull} which pulls the object to touch the fingers. E_{pull} also applies independently to each frame and we omit the superscript ^{n}. We here focus on the contact regions showcased in Fig 10. For each finger tip contact region with hand vertices $\{v_h\}_C$, the region-to-object distance is defined as the minimum distance of all (v_h, v_o) pairs. We use 5 finger tip regions and minimise the average of these region-to-object distances.

$$E_{pull} = \frac{1}{5} \sum_C d(\{v_h\}_C, V_o) \tag{11}$$

$$d(\{v_h\}_C, V_o) := \min_{v_h \in \{v_h\}_C, v_o \in V_o} \langle v_h - v_o, n_o \rangle \tag{12}$$

where n_o is the surface normal of v_o.

Pose initialisation for Static segments. As the object is typically supported by a surface when static, we use 10 initialisations all with an *upright* orientation. The initialisations differ in the object's rotation around the axis of support.

Pose initialisation for Stable Grasp segments. When using datasets with 3D ground truth, the initial rotations are generated by clustering the ground-truth rotations, where clustering is performed via the axis-angle representation of the rotation matrix. The initial translation is generated by averaging the ground-truth translations. We initialise 10 rotations and 1 global translation for each (object, left/right hand) pair. For EPIC-HIT, we manually set initial object relative poses to the common poses of each category. Each (category, left/right hand) pair has on average 4.1, minimum 1 and maximum 8 initialisation poses. Importantly, all compared methods (HOMan [29], Rigid, Dynamic, COP) are initialised with these same set of initial poses, ensuring fair comparison in Table-2 and Table-3 in the main paper and Table-3 in the supp.

Pose initialisation for Unstable Contact segments. We use random initialisation.

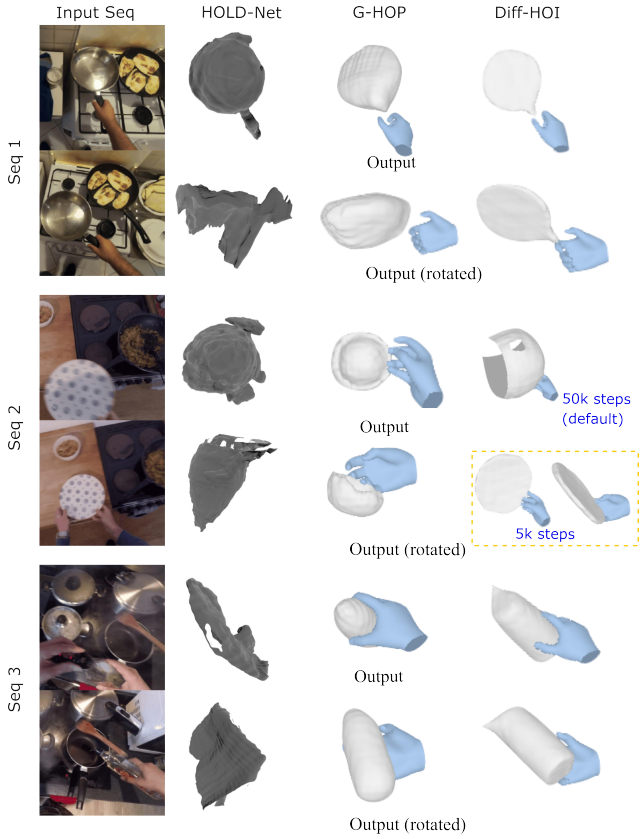

Figure 12. **In-the-wild qualitative evaluation of [21, 74, 75].** Owing to high occlusion due to fingers, the CAD-agnostic methods struggle to reconstruct the object shapes.

Computational cost analysis. The main computation is due to mesh projection in E_{mask}; E_{SG} is lightweight for meshes with $\approx$ 500 vertices. The Hand-Interaction Timeline (HIT) propagation is inherently sequential, potential speed-up can be gains through engineering the per-segment optimisation, e.g. multiple initialisation in the same segment can be optimised in parallel.

H. In-the-wild evaluation of CAD-Agnostic methods

In our method, we assume knowledge of the CAD model. We explore works that attempt reconstruction without CAD model's knowledge. In this section, we showcase these models to be unusable for in-the-wild hand-object reconstruction.

We evaluate CAD-agnostic methods HOLD-Net [21], G-HOP [75] and Diff-HOI [74] on the Stable Grasp from EPIC-HIT dataset. HOLD-Net is a neural rendering based multiple-view method, while G-HOP and Diff-HOI are data-driven methods that learn implicit shape priors from in-the-lab datasets.

Method	ADD↑	SCA-ADD↑	err_{rot} (°) ↓	err_{trans} (cm) ↓
FoundPose [48]	5.0	1.3	78.0	33.5
FoundPose [48] w/o texture	0.0	0.0	100.1	28.9
COP (Ours)	**70.0**	**32.9**	**39.2**	**1.3**

Table 13. Comparison with data-driven methods. We show avg. rotation and translation errors.

Figure 12 shows HOLD-Net is able to reconstruct the object's visible surface. However, HOLD-Net is unable to generate the complete object surface due to finger occlusion. As input views are typically limited in egocentric videos, HOLD-Net also struggles with the unseen surfaces – the bottle's symmetry is not reconstructed, see the rotated output. In-the-wild videos are also challenging for data-driven methods[2]. In Fig. 12, G-HOP fails to produce the shape for the pan and generates a bowl shape for the plate. Diff-HOI also performs poorly. Diff-HOI can generate the plate shape at an intermediate step (see 5K steps result in yellow square), but produces a wrong shape eventually (at the default 50k steps), highlighting robustness limitations.

Overall, these methods are at an infancy stage. Our method can extend to CAD-Agnostic methods when these are more robust. Importantly, it is not obvious how to quantitatively compare these methods on the same CAD-based metrics due to the need for alignment of the predicted shapes to the ground-truth CAD-model. This alignment is not obvious and has a significant impact on the numerical evaluation.

We also compare against FoundPose [48], a CAD-known but training free method. FoundPose use DINOv2 [47] to build correspondence between the image and the CAD model. Unlike other object pose estimator, FoundPose does not require training, therefore has the potential of scaling to unseen objects. In Tab. 13, we compare results on a random subset of 40 HOT3D stable grasp sequences. FoundPose is significantly worse than COP. Note that FoundPose relies on the texture of the instance CAD model, which is not a requirement for our method.

I. Limitations and Future Direction

Whilst results in-the-wild are very promising, our pipeline relies on hand pose estimation as a first stage. Despite the robustness incorporated by the multiple-view joint optimisation, our method fails when the predicted hand poses are incorrect (see main paper Figure 8). Our method also struggles with extreme occlusions and ambiguity from limited views.

Another limitation of our approach is its reliance on the knowledge of the category's CAD model. We show in Sec. H that current CAD-agnostic methods [21, 74, 75] struggle in-the-wild. CAD-agnostic reconstruction and generalisation to unknown objects is the ultimate goal, however current approaches do not provide sufficiently representative shapes for hand-object reconstruction where accurate object vertices are required for predicting contact.

[2]authors of these papers acknowledge their limitations in in-the-wild

Figure 13. SAM-3D results on EPIC-HIT frames from the Static segments of interaction timelines (i.e. when object is not in contact). The predicted models match the CAD models in EPIC-HIT and showcase potential extension into CAD-free assumption.

In addition, we note that the recently published SAM-3D model [9] could be used to obtain candidate CAD models, examples shown in Figure 13. While SAM3D is not integrated into the proposed pipeline, this is a plausible direction to address the known-CAD limitation.

Finally, we note that our definition of stable grasp is geometry-based. Exploring force closure and physical stability is left for future works.