GRASP-INFORMED OBJECT POSE ESTIMATION UNDER HAND OCCLUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object pose estimation is essential for extended reality to respond to users' interactions with physical objects. Yet hands naturally obscure parts of an object during interaction, making different poses look alike in egocentric images. We present HIROP, a method for estimating the 6-DoF pose of novel objects by jointly using visual evidence and grasp priors. HIROP uses a few examples of how an object is held to guide pose estimation without retraining the underlying model. At inference, these examples constrain how the object can be oriented relative to the observed hand and where the two should contact. Combined with visual information, these constraints help distinguish visually ambiguous poses. Experiments on HOT3D-Aria and DexYCB show improved pose accuracy over visual and hand–object interaction baselines across two visual implementations, including under severe hand occlusion. On HOT3D-Aria, using eight representative grasps per object and hand side, HIROP improves ADD-0.1d accuracy by 18.9 percentage points over the strongest compared baseline.

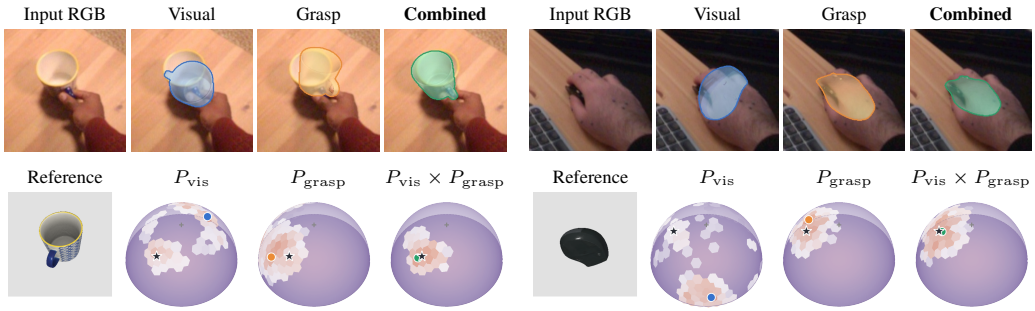


Figure 1: Jointly combining visual and grasp evidences constraints object pose ambiguity. Top: input image and pose selected from each cue. Below: reference and probability distribution over the rotation ($\star$: ground truth; colored dots: the overlaid poses with highest probabilities). Each case illustrates both visual and grasp ambiguity: Mug shows symmetry-induced rotational ambiguity due to handle occlusion and mouse shows ambiguity caused from severe occlusion by hand. Grasp alone is insufficient to finalize the pose due to its coarse nature. The product $P_{\text{vis}} \times P_{\text{grasp}}$ emphasizes shared support narrow to the true viewpoint.

1 INTRODUCTION

Object pose perception on eXtended Reality(XR) enables applications ranging from using a real pencil as a virtual magic wand to guiding furniture assembly with virtual overlay on physical objects. Hands are our primary interactor with the physical world. Yet the hands enabling these interactions also occlude the objects, making them difficult to track.

Model-based novel-object pose estimators use a known 3D model, supplied during onboarding, to render visual references for an object unseen during training (Örnek et al., 2024; Nguyen et al., 2024b; Labbé et al., 2022; Moon et al., 2025). Hand occlusion can make distinct poses visually indistinguishable, leaving ambiguity that training solely on object appearance alone cannot resolve (Manhardt et al., 2019; Nguyen et al., 2024a) (Fig. 1). Hand–object interaction (HOI) perception methods reduce such ambiguity by exploiting the knowledge of how an object is grasped during interaction. These methods use implicitly learned geometric features (Liu et al., 2021; Lin et al., 2023; Qi et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: How HOI perception methods use geometric and contact information. Integration stage shows when they are utilized; visual coupling describes how they are combined with image evidence.

Method	Geometric	Contact	Integration stage	Visual coupling
Liu et al., HFL-Net	Cross-attention	–	Pose estimation	Implicit
HOISDF	SDF weighting	–	Pose estimation	Implicit
CPF, ContactOpt	–	Contact refinement	Post optimization	None
ArtiBoost	Augmenting data	Grasp fitting	Training	Implicit
VPHO	Shared backbone	Plausibility prediction	Candidate selection	Sequential
HIROP(ours)	Grasp-relative rotation constraint	Depth initialization and correction	Proposal, refinement, selection	Joint objective

2024; Yang et al., 2022a) or predicted contact for post-hoc refinement (Yang et al., 2021; Grady et al., 2021) or as a secondary cue (Zhou et al., 2026). Instance-level HOI pose estimators learn diverse hand–object relations from annotated interactions involving the target objects.

We propose HIROP[1], a method for model-based novel-object pose estimation that is robust to hand–object interaction. A grasp prior consistent with the object's affordance helps constrain the object pose relative to the hand. Specifically, our joint objective seeks an object pose that explains the image and satisfies the geometric and contact relations of a compatible grasp prior. Offline onboarding augments mesh-derived visual templates with grasp priors from a small set of captured or generated interactions, without retraining the pose model. At inference, HIROP retrieves grasps with similar hand articulation and transfers their relations to the observed hand.

HIROP uses this joint objective to guide a staged proposal–refinement–selection pipeline, assigning grasp information a distinct role at each stage. The coarse stage generates multiple hypotheses supported by both cues, combining visual appearance scores with rotation votes from retrieved grasps. Refinement adjusts each hypothesis toward closer agreement by alternating visual pose updates and contact-conditioned depth corrections. Finally, selection evaluates the full joint energy at the refined hypotheses and returns the lowest-energy candidate.

Experiments on HOT3D-Aria (Banerjee et al., 2025) and DexYCB (Chao et al., 2021) demonstrate leading object pose accuracy against visual and HOI baselines, including on severely occluded frames under hand interaction. We show that pose accuracy remains stable as the grasp database is reduced to a small set of representative interactions. Ablation studies show the benefit of incorporating complementary grasp evidence throughout pose inference.

2 Related Work

Novel-object RGB pose estimation. Model-based methods estimate novel-object 6-DoF pose from RGB images through proposal, refinement, and selection. They prepare rendered templates from the known object mesh during offline onboarding and retrieve them for image matching at the proposal stage (Örnek et al., 2024; Nguyen et al., 2024b). MegaPose provides iterative render-and-compare refinement (Labbé et al., 2022), while Co-op and 3PT generate, refine, and score multiple visual hypotheses (Moon et al., 2025; Kalra et al., 2026). HIROP extends onboarding with representative grasp examples and transfers their hand–object relations to the observed hand for coarse-to-fine pose inference.

Hand–object interaction perception. HOI perception methods use hand–object geometry and contact to resolve pose ambiguity (Table 1). Liu et al. (2021), HFL-Net (Lin et al., 2023), and HOISDF (Qi et al., 2024) learn the geometric relations implicitly through feature cross-attention or signed-distance weighting. Contact-based refiners such as ContactOpt (Grady et al., 2021) and CPF (Yang et al., 2021) take initial pose estimates an input and adjust them toward predicted contact points without visual evidence, which keeps the object pose fixed or decreases its visual alignment. ArtiBoost (Yang et al., 2022a) augments training images with object-specific grasps synthesized through contact-constrained fitting. VPHO first aggregates visual candidates, then re-ranks them using force-based physical plausibility as a secondary cue (Zhou et al., 2026). HIROP only uses small representative grasp priors, retrieving geometric and contact relations and imposing them jointly with visual evidence.

[1]Hand Interaction Robust Object Perception.

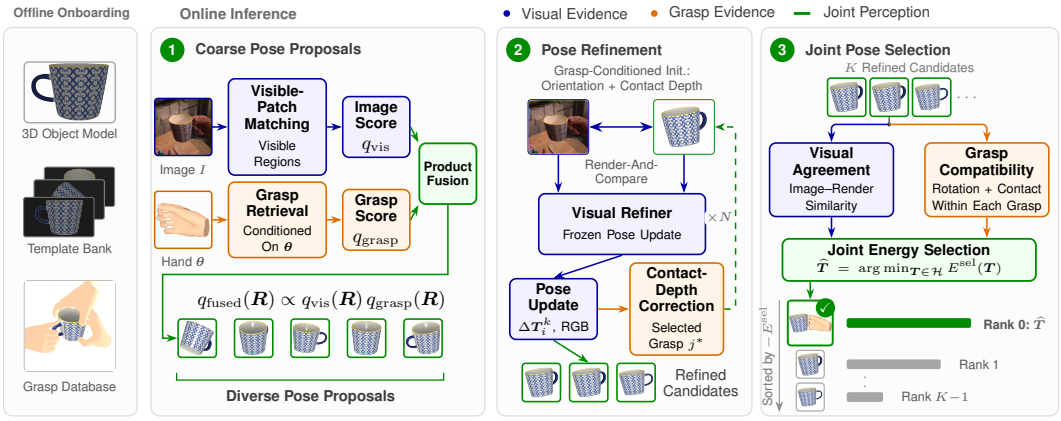


Figure 2: Overview: HIROP combines visual and grasp evidence throughout pose inference. The product of visual and grasp evidence emphasizes shared rotation support. Grasp-specific contact corrects depth during visual refinement. Selection evaluates both visual and orientation–contact support and chooses the lowest joint energy.

Hand grasp priors generation. These works model how object affordances and contact structure generate plausible hand grasps. OakInk and AffordPose provide affordance-aware hand configurations (Yang et al., 2022b; Jian et al., 2023). ContactGen (Liu et al., 2023) generates grasps through explicit contact modeling, while G-HOP (Ye et al., 2024) optimizes hand articulation and relative pose under a learned prior for a given object mesh. OpenHOI (Zhang et al., 2025) combines language-based affordance grounding with diffusion synthesis to support unseen objects. These methods could populate our grasp database without captured interactions.

3 Method

Overview. This section describes how HIROP constructs a grasp database and uses its relations for inference under a joint visual–grasp objective. A grasp database contains captured or generated examples of plausible grasps consistent with the object's affordances. Following established coarse-to-fine pose estimation pipelines (Örnek et al., 2024; Moon et al., 2025), grasp-informed proposal, refinement, and selection provide a staged approximation to this joint objective (Fig. 2).

Problem Definition. We seek an object pose that explains the observed image while remaining compatible with how the hand grasps the object. Given an RGB image $\boldsymbol{I}$ and the target object's 3D model, HIROP estimates its camera-frame 6-DoF pose $\boldsymbol{T} = (\boldsymbol{R}, \boldsymbol{t}) \in SE(3)$. A separately estimated hand pose provides MANO (Romero et al., 2017) articulation parameters $\boldsymbol{\theta} \in \mathbb{R}^{45}$ and a wrist coordinate frame $\boldsymbol{H}^{\text{obs}}$. Each object is onboarded with visual templates and a grasp prior database that is captured or generated.

3.1 Joint Visual–Grasp Objective

Our key idea is to use grasp examples as relational constraints, not as poses to reproduce. Figure 1 reveals that neither visual nor grasp evidence alone is sufficient to fully resolve the pose ambiguity. Visual cues can be ambiguous under occlusion, while grasp cues provide inherently coarse constraints that differ from the observed interaction. To favor poses supported by both cues, we define a joint energy that evaluates candidate poses using both visual and grasp evidence:

$$\boldsymbol{T}^{\star} = \arg \min_{\boldsymbol{T} \in SE(3)} \left[E_{\text{vis}}(\boldsymbol{T}) + E_{\text{grasp}}(\boldsymbol{T}) \right]. \tag{1}$$

E_{vis} measures the disagreement between the observed image and the appearance induced by the pose. E_{grasp} compares the relation between the observed hand and the candidate object pose with retrieved plausible hand–object relations. A hand–object relation consists of the object's hand-relative orientation and the contacting hand vertices. Relations from the retrieved examples are transferred to the observed hand, and their compatibility scores are aggregated as

$$E_{\text{grasp}}(\boldsymbol{T}) = -\log \sum_{j \in \mathcal{P}(\boldsymbol{\theta})} w_j \exp[-E_{\text{rotation},j}(\boldsymbol{R}) - E_{\text{contact},j}(\boldsymbol{T})]. \tag{2}$$

Here, $\mathcal{P}(\boldsymbol{\theta})$ contains the retrieved plausible grasps, and w_j are their normalized articulation-similarity weights (Sec. 3.3). Each mixture component evaluates the relative orientation and contact from the same grasp example, while the mixture allows different examples to support alternative poses. Because the joint objective can retain multiple pose modes, we search it by combining diverse proposals with local refinement, following established coarse-to-fine pipelines. The search comprises three stages: (i) proposal identifies promising poses using combined evidence; (ii) refinement adjusts each candidate using visual alignment and grasp-conditioned corrections; and (iii) selection compares the refined hypotheses under the full joint objective. The following sections describe how we build the grasp database and detail the stage-specific approximations of this common objective.

3.2 Grasp Database

Offline onboarding gathers representative grasp examples so that their geometric and contact relations remain available when visual evidence becomes ambiguous. Alongside visual templates rendered from the object mesh, we construct a grasp database from captured interactions or generated affordance-based configurations. A user can capture a few frames of grasps that match everyday usage (e.g., holding a mug or spatula by its handle) with the object's features clearly visible, allowing hand reconstruction and mesh-based object pose estimation to recover their relative configurations. Note that offline contact-based fitting can further refine these configurations since this capture does not need to be aligned with camera view appearance. Alternatively, affordance-conditioned hand generation can supply plausible examples directly from the object mesh (Sec. 2). Section 4.1 describes how we implement this construction process for benchmark evaluation.

Each database entry encodes transferable orientation and contact relations through the hand-relative object pose and the contacting hand vertices. For each object, the database is $\mathcal{D} = \{(\boldsymbol{\theta}_j, \boldsymbol{\Delta}_j, \mathcal{C}_j)\}_{j=1}^{N}$, where j indexes the N grasp examples. Each entry stores the hand articulation $\boldsymbol{\theta}_j$, the hand-relative object pose $\boldsymbol{\Delta}_j = (\boldsymbol{H}_j^{\mathrm{db}})^{-1}\boldsymbol{T}_j$, and the contact-vertex indices $\mathcal{C}_j$, where $\boldsymbol{H}_j^{\mathrm{db}}$ and $\boldsymbol{T}_j$ are the database hand frame and object pose in a common world coordinate system. Articulation similarity retrieves examples of how the observed hand may hold the object. We transfer $\boldsymbol{\Delta}_j$ through the observed hand frame $\boldsymbol{H}^{\mathrm{obs}}$, while the object translation is determined from the current visual and contact evidence. Contact-vertex indices are transferred to the observed hand mesh in frame $\boldsymbol{H}^{\mathrm{obs}}$, identifying hand locations likely to contact the object while allowing its pose to adapt to the current hand geometry.

3.3 Coarse Pose Proposals

Coarse proposals combine visual evidence with grasp rotation votes to retain object rotation candidates supported by both cues. Grasp information is inherently coarse: an object can be grasped in many ways, and a plausible database grasp need not exactly match the observed interaction. Directly imposing its stored translation and contact configuration can conflict with the actual grasp. However, the hand-relative object rotation provides necessary information for similar hand articulations, especially when visual evidence is ambiguous. We approximate grasp compatibility in the coarse stage using retrieved rotation information, as explained below.

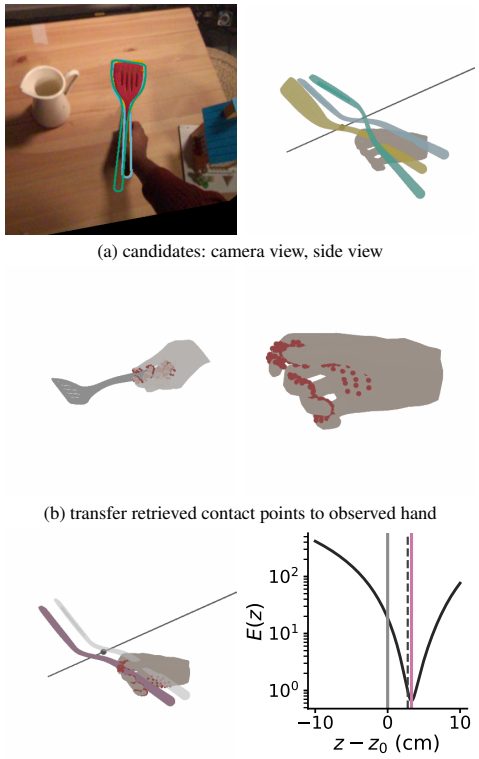
Grasp orientation compatibility. Retrieved grasps must retain alternatives rather than impose a fixed rotation, since similar hand articulations can support several object rotations. We first retrieve the k nearest examples of the observed hand side in articulation space and assign them normalized Gaussian similarity weights w_j (Appendix A). Each retrieved grasp provides a reference rotation $\boldsymbol{R}_j$ by transferring its hand-relative object rotation to the observed hand. To allow small deviations from the transferred rotation, we penalize angular deviation from each reference rotation using the isotropic matrix-Fisher penalty $E_{\mathrm{rotation},j}(\boldsymbol{R}) = \kappa[3 - \mathrm{tr}(\boldsymbol{R}_j^{\top}\boldsymbol{R})]$, which is zero at $\boldsymbol{R}_j$ and increases with angular deviation; κ controls how tightly nearby rotations are constrained. For each candidate rotation $\boldsymbol{R}_g$ on a shared grid $\mathcal{G} = \{\boldsymbol{R}_g\}_{g=1}^{G}$, we sum the scores of the retrieved grasps, weighted by their articulation similarity, as $\sum_j w_j \exp[-E_{\mathrm{rotation},j}(\boldsymbol{R}_g)]$. We normalize these scores over all candidate rotations and add a small uniform component to obtain the probability distribution q_{grasp}. The uniform component allows rotations poorly represented in the grasp database to remain selectable when supported by strong visual evidence; Appendix A provides implementation details. The resulting distribution can support multiple plausible orientations (Fig. 1).

Visual orientation alignment. The visible part of the object identifies rotations consistent with the observed appearance. Following existing visual coarse-proposal methods, we render the object at each grid rotation $\boldsymbol{R}_g$ and encode its patch features. At inference, an object-centered crop keeps the query patches inside the object mask. Cosine similarity to each template, followed by a temperature softmax, gives $q_{\text{vis}}(\boldsymbol{R}_g)$.

Joint proposals. We combine visual and grasp evidence by multiplying their probabilities for each candidate rotation and normalizing over the shared rotation grid:

$$q_{\text{fused}}(\boldsymbol{R}_g) = \frac{q_{\text{vis}}(\boldsymbol{R}_g)\, q_{\text{grasp}}(\boldsymbol{R}_g)}{\sum_{h=1}^{G} q_{\text{vis}}(\boldsymbol{R}_h)\, q_{\text{grasp}}(\boldsymbol{R}_h)}. \quad (3)$$

When several rotations receive similar visual scores, grasp compatibility can distinguish them. Conversely, when the grasp supports several rotations similarly well, visual evidence can distinguish them by their appearance (Fig. 1). Because refinement adjusts each starting pose only locally, we retain $K = 5$ high-scoring rotations with a minimum angular separation between them, keeping distinct alternatives for refinement. When both hands contact the object, we generate coarse candidates from each hand rather than choosing only one, and select the best pose from their combined set.



(a) candidates: camera view, side view

(b) transfer retrieved contact points to observed hand

(c) depth correction, energy along the ray

Figure 3: Contact-constrained refinement, on one frame of a grasped spatula. (a) depth-ambiguous candidates through visual evidence; (b) the retrieved contact region and transferred vertices; (c) depth correction example with the energy along the ray: refiner depth (grey), corrected (purple), GT (dashed).

3.4 Contact-Constrained Refinement

Refinement complements visual pose updates with contact-based depth correction. Under occlusion, visually similar candidates can lie centimeters apart along the viewing ray, and contact relations help distinguish their depths (Fig. 3(a)). For each proposal, we use the contacts of one grasp retrieved by hand articulation similarity. The visual refiner updates the full pose, and each subsequent contact correction adjusts only the depth.

Specifically, we choose the retrieved grasp j^* whose transferred object rotation is closest to the proposal rotation $\boldsymbol{R}_i$. Its contact-vertex indices $\mathcal{C}_{j^*}$ are mapped to contact locations on the current hand using the observed articulation $\boldsymbol{\theta}$ and hand frame $\boldsymbol{H}^{\text{obs}}$ (Fig. 3(b)). We initialize the depth by moving the object along the viewing ray to minimize the mean distance from these hand vertices to the object surface. The contact penalty $E_{\text{contact},j}$ measures the separation between grasp j's transferred hand vertices and the posed object surface, allowing a small gap δ; Appendix A gives its exact form.

Alternating visual updates and contact corrections improves image alignment while anchoring the depth to the hand. The RGB visual refiner iteratively proposes a local pose update from the observation and a render at the current pose. Between its iterations, with the updated rotation fixed, we correct the depth by minimizing

$$E_{\text{depth}}(z) = \frac{(z - z_0)^2}{2\sigma^2} + E_{\text{contact},j^*}\big(\boldsymbol{R}_i, \boldsymbol{t}_i(z)\big), \quad (4)$$

where z_0 is the refiner's depth and $\boldsymbol{t}_i(z)$ moves the translation along the viewing ray (Fig. 3(c)). The first term discourages departures from the visual depth estimate, with σ setting the distance scale; the second encourages the object surface to remain near the transferred hand contacts. A final visual update yields each refined pose $\widetilde{\boldsymbol{T}}_i$.

3.5 JOINT POSE SELECTION

The refined poses $\mathcal{H} = \{\widetilde{\boldsymbol{T}}_i\}_{i=1}^{K}$ are evaluated, and the pose with the lowest combined visual–grasp energy is selected:

$$\widehat{\boldsymbol{T}} = \arg\min_{\boldsymbol{T}\in\mathcal{H}} \left[E_{\text{vis}}^{\text{sel}}(\boldsymbol{T}) + \tau E_{\text{grasp}}^{\text{sel}}(\boldsymbol{T})\right], \quad \text{where} \quad E_{\text{vis}}^{\text{sel}}(\boldsymbol{T}) = -a(\boldsymbol{T}) - \eta \log p(\boldsymbol{T}),$$
$$E_{\text{grasp}}^{\text{sel}}(\boldsymbol{T}) = -\log \sum_j w_j \exp[-\alpha_R E_{\text{rotation},j}(\boldsymbol{R}) - \alpha_C E_{\text{contact},j}(\boldsymbol{T})]. \tag{5}$$

Here, $\tau > 0$ balances grasp compatibility against visual agreement, η weights the visual model's confidence, and α_R and α_C weight the rotation and contact penalties.

Visual agreement builds on the hand–object depth relations established during contact-constrained refinement. We render each refined object pose together with the observed hand and compare the render with the observed image only at patches where the object remains visible. The mean DINOv2 similarity over these patches, $a(\boldsymbol{T})$, improves the visual refiner's confidence score $p(\boldsymbol{T})$ in terms of appearance matching evaluation.

Grasp compatibility recomputes the orientation and contact penalties of each refined pose against all retrieved grasps. Each grasp contributes according to its articulation similarity and its combined rotation and contact penalties. A reduced orientation weight accommodates deviations from the stored rotation votes. These scoring constants are tuned on training frames from a single dataset and fixed for different datasets or our variations (Appendix A).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Protocol. HOT3D-Aria (Banerjee et al., 2025) includes daily interaction scenarios captured by egocentric XR. We evaluate object pose on interaction frames only: with a minimum hand–object surface distance of at most 1 cm. We split two subjects out of the training set since official test annotations are withheld and sampled with 2 FPS, resulting 4,110 interaction frames on 33 objects. The other seven subjects interaction data provide the training split for instance-level HOI baselines. We annotate ground truth object poses for 292 additional severely occluded frames ground truths are not provided, yielding 4,402 evaluation samples. The annotation procedure is described in Appendix A. To evaluate performance under severe occlusion, we select samples at or below the 10th visibility percentile for each object, resulting in 454 highly occluded frames. DexYCB (Chao et al., 2021) captures object pick-and-place interactions. We evaluate the S0 test split, using the single viewframe behind the participant's shoulder to approximate an egocentric view with high occlusion. We evaluate all baselines on the same definition of interaction frames, resulting in 7,298 interaction frames on 20 objects. HIROP leverages SOTA RGB-based hand prediction model (Potamias et al., 2025) with Procrustes-Aligned coordinate system using wrist position as GT, same as the HOI baseline settings.

Grasp onboarding. We evaluate two different onboarding method: user-captured and generated. We approximate user capture process by manually selecting a small subset of annotated interactions from training splits of both datasets. We pick 8 representative frames per hand, showing its most natural everyday grasp for each object, such as holding a kettle or a spatula by its handle. This makes our approximated captured database a subset of the training dataset at a 0.27% scale. For generated onboarding, we use a G-Hop (Ye et al., 2024) to produce 8 diverse right-hand grasps per object. Since G-Hop does not support unseen objects, we evaluate this on DexYCB dataset only. Manual screening of generated objects is needed since G-Hop does not always produce plausible grasps.

Baselines. Model-based novel-object baselines without hand input are FoundPose (Örnek et al., 2024) with MegaPose refinement (Labbé et al., 2022), GigaPose (Nguyen et al., 2024b), and Co-op (Moon et al., 2025). We train Co-op ourselves, reaching 93% and 97% of its published AR on LM-O and YCB-V. Each method refines 5 hypotheses per object as our single-hand pipeline. Instance-level, seen-object HOI baselines are HFL-Net (Lin et al., 2023), HOISDF (Qi et al., 2024), and VPHO (Zhou et al., 2026), each supervised on the corresponding training split.

Table 2: Pose accuracy on HOT3D-Aria interaction frames. Best in bold, second-best underlined.

	Method	All (4,402)			Highly occluded (454)		
		$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$
Novel-object RGB pose (no hand input)	GigaPose	48.0	32.8	36.5	35.6	20.7	28.3
	FoundPose	63.0	47.1	51.5	49.8	33.0	41.9
	Co-op	65.7	53.7	57.2	53.8	43.6	46.8
HOI perception (seen-object)	HOISDF	46.3	14.0	19.9	39.4	15.4	17.1
	HFL-Net	64.1	37.1	53.6	49.1	23.6	41.6
	VPHO	76.8	50.0	60.6	68.8	39.6	50.8
	HIROP-f (ours)	<u>81.0</u>	<u>67.7</u>	<u>69.9</u>	<u>69.3</u>	<u>51.1</u>	<u>57.8</u>
	HIROP-c (ours)	**82.7**	**72.6**	**75.2**	**70.2**	**56.4**	**61.0**

Table 3: DexYCB interaction frames (S0 test, one near-egocentric camera, 7,298 frames, 20 objects). Best in bold, second-best underlined. †: supervised on the S0 training split; HIROP trains nothing.

	Method	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$
Novel-object RGB pose	GigaPose	43.2	47.0
	FoundPose	41.6	45.8
	Co-op	42.1	48.0
HOI perception	HOISDF†	28.3	26.5
	HFL-Net†	22.3	36.0
	VPHO†	51.3	50.2
	HIROP-f	**55.5**	**54.2**
	HIROP-c	<u>53.3</u>	<u>54.1</u>

Table 4: Onboarding source on DexYCB right-hand frames (3,732), since the generated database covers right hands only. Best in bold.

	DB	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$
HOISDF†	–	26.8	21.2
VPHO†	–	43.0	41.6
HIROP-f	Capt.	**50.1**	**48.1**
	Gen.	47.6	46.8
HIROP-c	Capt.	45.3	46.3
	Gen.	44.2	45.3

Metrics. We report average ADD-0.1d and its average recall over thresholds $\{0.05, \ldots, 0.50\}d$, denoted $\mathrm{AR}_{\mathrm{ADD}}$. ADD measures mean corresponding-vertex displacement; objects annotated as symmetric use nearest-vertex matching (ADD-S) instead. ADD-0.1d evaluates performance under harsh conditions, whereas $\mathrm{AR}_{\mathrm{ADD}}$ measures general accuracy. We also report metrics from BOP challenge (Hodaň et al., 2020), denoted $\mathrm{AR}_{\mathrm{BOP}}$: the mean VSD, MSSD, and MSPD recall on DexYCB, and mean MSSD and MSPD recall on HOT3D-Aria. These metrics assess visible-surface discrepancy and symmetry-aware maximum surface distance in 3D and in projection, respectively. All methods use the same evaluation samples, scorer, and official evaluation meshes.

4.2 Object Pose Accuracy on Interaction Frames

To test our grasp-based design across visual implementations, we compare two HIROP versions with the same grasp-guided proposal–refinement–selection pipeline. HIROP-f uses FoundPose-style template matching and MegaPose refinement; HIROP-c uses Co-op for proposal and refinement. On HOT3D-Aria, both HIROP versions outperform the compared baselines on both evaluation subsets (Table 2). HIROP-c exceeds the strongest baseline in ADD-0.1d by 18.9 points overall and 12.8 points under heavy occlusion. These results support the use of grasp evidence across both visual implementations. On DexYCB, HIROP-f achieves 55.5 ADD-0.1d and 54.2 $\mathrm{AR}_{\mathrm{BOP}}$, exceeding VPHO by 4.2 and 4.0 points, respectively (Table 3). HIROP-c also outperforms the compared baselines on both metrics. Table 4 shows the accuracy using the generated grasp prior database compared with the selected database. It outperforms the baselines, although accuracy is lower compared with the captured dataset.

4.3 Qualitative Results

Figure 4 compares estimated hand–object configurations in camera and side views. VPHO uses contact force as a secondary cue and produces plausible hand–object configurations, but its object estimates deviate from the visual evidence. Co-op is unaware of hand interaction; estimated poses suffer from rotation and depth errors stem from visual ambiguity of input images. HIROP more closely matches the ground-truth object by combining visual and grasp evidence.

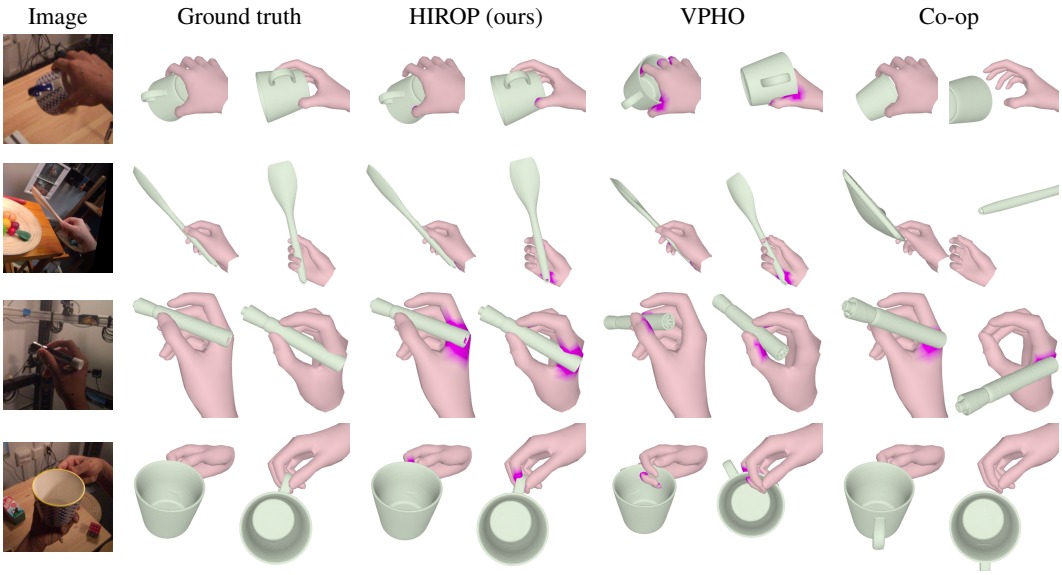


Figure 4: HOT3D-Aria interaction frames. Left: input crop. Right: ground truth and estimates from HIROP, VPHO, and Co-op in camera (left) and side (right) views. Pink marks penetration/contact.

Table 5: Stage-wise comparisons on HOT3D-Aria. F/H: FoundPose/HIROP proposal; M: MegaPose; C: contact-depth correction. Selection uses coarse (C), M, or H scores. Full results: Table 8.

				All		Highly occluded	
	Prop.	Ref.	Sel.	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d
1	F	—	C	49.3	17.9	42.4	13.0
2	H	—	C	57.3	19.9	47.9	12.1
3	H	M	M	74.8	56.7	61.7	39.0
4	H	M+C	M	78.2	61.9	66.0	45.2
5	H	M+C	H	**81.0**	**67.7**	**69.3**	**51.1**

Table 6: Design choices on HOT3D-Aria (All). (a) Rotation-only evaluation with GT translation; (b) depth-correction cues with shared proposals, MegaPose refinement, and MegaPose-confidence selection. Best in bold; full results in Table 9.

	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d
(a) Proposal		
Visual	60.8	37.1
+ prior grasp (ours)	66.4	38.1
+ prior grasp & contact	64.2	36.1
+ predicted plausibility	66.5	40.0
(b) Refinement: depth correction		
Prior position	76.0	56.1
Predicted contact	77.4	60.0
Predicted plausibility	62.4	48.9
Prior contact (ours)	**78.2**	**61.9**
Prior contact + rotation	77.8	61.5

Table 7: Hand input on HOT3D-Aria. Missed detections count as failures.

Hand	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$
Prediction (WiLoR)	82.7	72.6	75.2
Ground Truth	87.3	78.5	79.6

4.4 Ablation Studies

We conduct two ablation studies: examining the contribution of each pipeline stage (Table 5) and the design choices of grasp-aware modules within proposal and refinement (Table 6).

4.4.1 Stage Contributions

Proposal. We compare visual-only and fused visual–grasp proposals with the same initial object position (rows 1–2). Grasp evidence improves $\mathrm{AR}_{\mathrm{ADD}}$ by 8.0 points on all frames and 5.5 points on highly occluded frames. **Refinement.** Applying MegaPose to the grasp-guided proposals substantially improves pose accuracy (rows 2–3). We then initialize object depth using contact regions transferred from retrieved grasps and reapply contact-based depth correction between the five MegaPose iterations (row 4). With MegaPose-confidence selection unchanged, contact-based correction further improves ADD-0.1d by 5.2 points on all frames and 6.2 points on highly occluded frames. **Selection.** Keeping the refined candidates unchanged, we replace MegaPose-confidence selection with HIROP selection, which considers both visual agreement and grasp compatibility (rows 4–5). HIROP selection improves ADD-0.1d by 5.8 points on all frames and 5.9 points on highly occluded frames. The larger gain under occlusion supports using combined visual–grasp selection for final selection. To evaluate our design choices, we compare interaction cues for proposal and vary both

the interaction cue and the corrected pose components during refinement (Table 6; highly occluded results in Table 9).

4.4.2 Grasp-aware Design Choices

(a) Proposal. We evaluate rotation accuracy using ground-truth translation, with the visual-only variant using FoundPose-style proposals while ours combines visual evidence with the grasp-orientation prior. We also test adding prior-contact agreement to the fused ranking and replacing the grasp-orientation prior with VPHO's predicted physical plausibility (Zhou et al., 2026); implementation details are provided in Appendix A. Grasp-conditioned proposals improve $\mathrm{AR}_{\mathrm{ADD}}$ over the FoundPose-style visual-only baseline by 5.6 points on all frames. Adding prior-contact agreement to proposal ranking reduces $\mathrm{AR}_{\mathrm{ADD}}$ by 2.2 points. The grasp prior achieves performance comparable to predicted plausibility. These results support using grasp-orientation evidence to guide object rotations at the proposal stage. **(b) Refinement.** All variants use the same proposals, five MegaPose iterations, and MegaPose-confidence selection. We correct depth using retrieved grasp positions, CHOI-predicted hand contacts (Hu et al., 2024), or VPHO's physics-based score. Appendix A details each correction and its application at initialization and between refinement iterations. Prior-contact depth correction achieves 61.9 ADD-0.1d on all frames, compared with 60.0 for predicted contact and 56.1 for prior-position correction. Keeping our prior-contact cue unchanged, we compare correcting depth alone with correcting both depth and rotation. In the depth-only variant, rotation remains governed by the visual refinement updates. Depth-only correction achieves ADD-0.1d of 61.9 versus 61.5 for joint depth and rotation correction on all frames, and 45.2 versus 42.7 on highly occluded frames. Together, these ablations support using grasp-orientation evidence to guide proposals and prior contact to complement visual refinement through depth correction.

4.5 Analysis

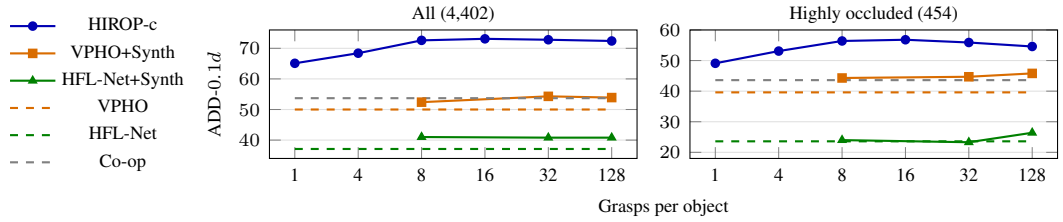


Figure 5: Grasp database size on HOT3D-Aria. +Synth: fine-tuned with augmentation datas synthesized based on the identical grasp frames from database. Horizontal lines use no grasp data.

Grasp Database Size. HIROP's accuracy largely saturates at eight grasps per object and hand side (Figure 5). We vary the grasp count using farthest-point sampling order in every interaction frame captured from the training set. We also fine-tune VPHO and HFL-Net with synthetic hand-object images generated from the corresponding grasp sets using ArtiBoost (Yang et al., 2022a). Despite this augmentation, their accuracy remains substantially below ours. The limited gains beyond eight grasps suggest that how well the grasp priors represent the interactions matters more than their quantity. **Hand Inputs.** To assess the effect of hand input quality, we replace the WiLoR-predicted hands used in our main experiments with ground-truth hands (Table 7). These gains suggest further room for improvement through more accurate and reliable hand inputs.

5 Conclusion

HIROP uses the grasping hand to constrain object pose when visual evidence is occluded. It treats grasp examples as transferable orientation and contact relations, coupling them with visual alignment in a joint objective. Coarse proposals combine visual and grasp orientation evidence, refinement corrects depth through grasp-specific contact, and selection evaluates the refined poses under joint visual–grasp compatibility. Experiments on HOT3D-Aria and DexYCB show that the full pipeline improves pose accuracy over visual baselines, including under severe hand occlusion.

Potential Limitations. Under the extreme low visibility cases, visual refinement can move a plausible pose away from the correct solution. An important extension is to detect this status and exclude object visual evidence, constraining pose through the grasp prior and temporal information until informative observations return.

AI USE STATEMENT

In this work, we used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, and assist with translation. We have not used generative AI tools to support qualitative and thematic data analysis and interpret results, and generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, edit a research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. The authors have checked LLM-generated translated output writing for clarity. LLM-guided related literature search was verified by comparing with actual literatures manually. Implemented code was verified and tested for correctness by first and second authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marc Alexa. Super-Fibonacci spirals: Fast, low-discrepancy sampling of SO(3). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomáš Hodaň. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7061–7071, 2025.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

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 (ECCVW), 2020.

Junxing Hu, Hongwen Zhang, Zerui Chen, Mengcheng Li, Yunlong Wang, Yebin Liu, and Zhenan Sun. Learning explicit contact for implicit reconstruction of hand-held objects from monocular images. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 38, pp. 2220–2228, 2024.

Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. AffordPose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14713–14724, 2023.

Agastya Kalra, Tim Salzmann, Guy Stoppi, Dmitrii Marin, Rishav Agarwal, Vage Taamazyan, Martin Bokeloh, Stefan Hinterstoisser, Anton Boykov, Alberto Dall'Olio, Pravin Dangol, Kartik Venkataraman, and Huaijin Chen. 3D-object perception transformer (3PT). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25777–25787, 2026.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D pose estimation of novel objects via render & compare. In Proceedings of the Conference on Robot Learning (CoRL), 2022.

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 (CVPR), 2023.

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 (CVPR), pp. 14687–14697, 2021.

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. ContactGen: Generative contact modeling for grasp generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6D pose from visual data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Co-op: Correspondence-based novel object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. NOPE: Novel object pose estimation from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024a.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and robust novel object pose estimation via one correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024b.

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 (ECCV), 2024.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12242–12254, 2025.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. HOISDF: Constraining 3D hand-object pose estimation with global signed distance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10392–10402, 2024.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG), 36(6):245:1–245:17, 2017.

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/CVF International Conference on Computer Vision (ICCV), 2021.

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 (CVPR), 2022a.

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 (CVPR), pp. 20953–20962, 2022b.

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

Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. OpenHOI: Open-world hand-object interaction synthesis with multimodal large language model. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025.

Jun Zhou, Chi Xu, Kaifeng Tang, Yuting Ge, Tingrui Guo, and Li Cheng. VPHO: Joint visual-physical cue learning and aggregation for hand-object pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), volume 40, pp. 13683–13691, 2026.

A Implementation Details

Additional pose annotations. For the 292 additional HOT3D-Aria frames, we manually adjust each object's 3D translation and rotation to align its projected mesh with the RGB image, referring to neighboring frames.

Grasp retrieval and hypothesis generation. We retrieve $k = 16$ neighbors in MANO's 45-dimensional PCA articulation space, using $\tau_{\text{art}} = 1$ for the similarity weights.

$$w_j = \frac{\exp\big(-\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2^2 / 2\tau_{\text{art}}^2\big)}{\sum_{l \in \mathcal{N}(\boldsymbol{\theta})} \exp(-\|\boldsymbol{\theta} - \boldsymbol{\theta}_l\|_2^2 / 2\tau_{\text{art}}^2)}. \tag{6}$$

We evaluate the rotation mixture on $\mathcal{G}$ as

$$q_{\text{grasp}}(\boldsymbol{R}_g) = (1 - \epsilon)\frac{1}{Z}\sum_j w_j \exp[-E_{\text{rotation},j}(\boldsymbol{R}_g)] + \frac{\epsilon}{G}, \tag{7}$$

where Z normalizes the kernel mixture over the grid.

The orientation prior uses matrix-Fisher kernels with a 15° bandwidth and uniform weight $\epsilon = 0.01$. The shared grid contains $G = 4096$ super-Fibonacci rotations (Alexa, 2022). We select $K = 5$ fused hypotheses with at least 45° pairwise geodesic separation. For objects held by both hands, candidates are generated and refined per hand, then pooled for final selection.

Visual templates. We use DINOv2 ViT-L/14 with registers, extracting layer-18 patch tokens from 420×420 renders. The object's bounding-box longer side occupies 0.6 of the crop. Descriptors are reduced to 256 dimensions by PCA and encoded with a 2,048-word vocabulary, 3-NN soft assignment ($\sigma = 10$), and tf–idf weighting. The query uses FoundPose's padded virtual-camera crop. Candidate renders use crop side length $f\, d/(0.6z)$ for object diameter d and depth z.

Depth initialization. The initial viewing ray $\boldsymbol{r}$ follows the crop camera's optical axis, defined by the mean of the box corners' unit rays and normalized to unit camera-frame z. For mesh bounding-box center $\boldsymbol{c}$ and fixed rotation $\boldsymbol{R}_i$, depth z gives $\boldsymbol{t}_i(z) = z\boldsymbol{r} - \boldsymbol{R}_i\boldsymbol{c}$. An apparent-size estimate $z_{\text{box}} = f_x d/\ell$, with ℓ the longer box side, centers the depth search. We evaluate mean contact-vertex distance on a 2 mm grid over $z_{\text{box}} \pm 25$ cm and retain its minimizer. If no grasp is retrieved, depth defaults to the mean of z_{box} and the camera-frame depth of the palm root (MANO joint 0).

Contact-constrained refinement. For grasp j, let $\rho_m(\boldsymbol{T})$ be the shortest distance from observed hand vertex $m \in \mathcal{C}_j$ to the object surface at pose $\boldsymbol{T}$. The contact penalty averages squared gaps exceeding a tolerance δ:

$$E_{\text{contact},j}(\boldsymbol{T}) = \frac{1}{|\mathcal{C}_j|}\sum_{m \in \mathcal{C}_j}\left(\frac{\max(\rho_m(\boldsymbol{T}) - \delta,\, 0)}{\nu}\right)^2, \tag{8}$$

where ν sets the distance scale. Vertices within δ of the surface incur no penalty; the nearest object-surface point is recomputed as the candidate pose changes. Each hypothesis receives five RGB refinement iterations from the selected visual backbone: MegaPose for HIROP-f and Co-op for HIROP-c. After each of the first four iterations, we minimize Eq. 4 by Gauss–Newton with backtracking, using at most 20 steps, a 0.1 mm tolerance, and a 25 cm displacement cap. We set $\delta = 8$ mm, $\nu = 5$ mm, and $\sigma = 5$ cm. These values were fixed using a 150-frame calibration before evaluation. The fifth visual update is returned without a subsequent depth correction.

Selection. The image–render comparison a_i is the mean of a 16×16 DINOv2 cosine-similarity map over the cells where the rendered object is visible; the render uses the query crop's camera so the two are pixel-aligned. p_i is the refiner's per-hypothesis match probability: MegaPose returns it directly and Co-op returns its logit, both trained with a binary cross-entropy. At selection, Eq. 5 uses $\eta = 0.02$, $\tau = 0.05$, $\alpha_R = 0.03$, and $\alpha_C = 0.3$, chosen once by a grid on the HOT3D training split and fixed for every result in this paper, including DexYCB and the Co-op variant; the top of the grid is flat to 0.2 points of selection recall. Using $\alpha_R = 1$ instead costs 5 points, and with the two grasp terms marginalized over the examples independently rather than jointly, selection recall changes by under 0.3 points.

Co-op integration. For HIROP-c, Co-op provides the visual evidence for proposal and refinement, and its pose-selector confidence enters our joint final selection. At proposal, we map Co-op's 42 template-view scores to the shared grid $\mathcal{G}$ by assigning each grid rotation the score of the nearest template viewpoint. We normalize the lifted scores to q_{vis} and combine them with q_{grasp} using Eq. 3. We select K fused modes corresponding to distinct Co-op viewpoints and initialize a 6DoF candidate for each mode with the correspondence-based PnP estimate associated with its template. For refinement, we initialize candidate depth from the transferred contacts and apply the correction in Eq. 4 after each of the first four Co-op refinement updates; the fifth update is retained without correction. For selection, we convert each Co-op pose-selector logit with a sigmoid to the match probability $p(\boldsymbol{T})$ in Eq. 5. The final joint energy combines this probability with the DINOv2 image–render agreement $a(\boldsymbol{T})$ and the same coupled rotation–contact grasp term used for HIROP-f.

A.1 Ablation Implementation Details

We describe the alternatives in Table 6: interaction cues for ranking proposals and correcting poses during refinement.

A.1.1 Proposal

The visual-only variant uses FoundPose-style proposals, while our proposal combines visual evidence with the grasp-orientation prior. We evaluate the top-ranked rotation from each variant with ground-truth object translation substituted only for evaluation.

Prior-contact re-ranking. The visual + prior grasp & contact variant re-ranks the 16 highest-scoring, rotation-diverse fused candidates using a contact-misfit term. For each candidate rotation, we select the retrieved grasp with the closest hand-relative object orientation and apply its contact mask to the corresponding vertices of the current hand mesh. Keeping the candidate rotation fixed, we initialize depth by a contact-distance search and then fit translation by nearest-surface correspondences. The mean contact-to-surface distance after this fit is the contact misfit; the fitted translation is not used in the rotation-only evaluation. We divide the fused score and contact misfit by their respective standard deviations over the candidate pool, and subtract the normalized misfit from the normalized fused score with equal weight.

Predicted plausibility and score fusion. We replace the grasp-orientation prior with the physics-based score of VPHO (Zhou et al., 2026). For this proposal, the predicted local forces are expressed at force anchors on the WiLoR hand mesh. Following VPHO's local-to-global force transformation, we reconstruct the anchor positions and local frames on the WiLoR hand mesh and rotate the predicted local forces accordingly. We evaluate the anchors, forces, and object poses in a common camera coordinate system, accounting for VPHO's left-hand reflection convention. Each candidate rotation is scored at the fixed contact-free object-center initialization. Let $\mathbf{a}_i$ and $\mathbf{f}_i$ denote a force anchor and its predicted force in camera coordinates, $\mathbf{b}_i(T)$ the nearest transformed object vertex, and $\mathbf{c}(T)$ the transformed object center of mass. The implemented score is the product of a force-weighted distance and a moment term:

$$D(T) = \sum_i \frac{\|\mathbf{f}_i\|}{\sum_j \|\mathbf{f}_j\|} \|\mathbf{a}_i - \mathbf{b}_i(T)\|, \tag{9}$$

$$L(T) = \left\| \sum_i \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|} \times \big(\mathbf{a}_i - \mathbf{b}_i(T) - \mathbf{c}(T)\big) \right\|, \qquad E_{\text{phys}}(T) = D(T)L(T). \tag{10}$$

This expression retains the moment convention of the VPHO scoring implementation. To combine the two signals, we center the visual log scores by their median and divide by their interquartile range. We rank the physics energies in ascending order, map their ranks to $[0, 1]$, and subtract this rank penalty from the normalized visual score with weight 1. Candidate rotations use the same rotation grid and diversity selection as the grasp-guided proposal.

A.1.2 Refinement

Shared candidates and update schedule. Using the Aria-training DB-8, all variants start from the same five proposals per hand; candidates from both hands are pooled for MegaPose-confidence

Table 8: Full stage-wise comparisons using HIROP-f on HOT3D-Aria. Rows 1–3 share contact-free depth initialization; row 4 adds contact-based initialization and iterative depth correction. Rows 4–5 select from identical refined candidates.

				All (4,402)			Highly occluded (454)		
	Proposal	Refinement	Selection	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$
1	FoundPose	—	coarse score	49.3	17.9	30.8	42.4	13.0	24.0
2	HIROP	—	coarse score	57.3	19.9	37.4	47.9	12.1	30.6
3	HIROP	MegaPose	MegaPose score	74.8	56.7	63.2	61.7	39.0	50.8
4	HIROP	+ contact depth	MegaPose score	78.2	61.9	65.9	66.0	45.2	53.4
5	HIROP	+ contact depth	HIROP selector	**81.0**	**67.7**	**69.9**	**69.3**	**51.1**	**57.8**

selection. Corrections are applied at initialization and between five MegaPose updates, with no correction after the last update. The evaluation uses 4,402 samples for the full set and 454 for the highly occluded subset; missing estimates count as failures.

Prior position. We use the articulation-based grasp retrieval and similarity weights described above. The retrieved grasps' wrist-relative object translations are transferred to the current hand frame, transformed to camera coordinates, and averaged with these weights to obtain $\mathbf{t}_{\mathrm{prior}}$. For candidate rotation R, translation $\mathbf{t}$, and object-space bounding-box center $\mathbf{c}_{\mathrm{box}}$, we define the unit viewing ray and corrected ray distance as

$$\widehat{\mathbf{r}} = \frac{\mathbf{t} + R\mathbf{c}_{\mathrm{box}}}{\|\mathbf{t} + R\mathbf{c}_{\mathrm{box}}\|}, \qquad s = (\mathbf{t}_{\mathrm{prior}} + R\mathbf{c}_{\mathrm{box}})^{\top}\widehat{\mathbf{r}}. \tag{11}$$

The corrected translation is $s\widehat{\mathbf{r}} - R\mathbf{c}_{\mathrm{box}}$. This projects the database-derived position onto the candidate's viewing ray while leaving its rotation unchanged.

Predicted contact. CHOI (Hu et al., 2024) predicts per-vertex contact probabilities from the WiLoR-based crop. We match vertices with probability above 0.5 to their nearest projected object-surface points and use the median depth offset to translate the object along its viewing ray, keeping rotation fixed. We require at least three matches within 30 pixels, ignore offsets within 1 cm, and cap each depth update at 5 cm.

Predicted plausibility. We use the hand estimate obtained after VPHO's visual hand aggregation and its corresponding predicted forces, as used in the subsequent object-physics scoring stage. The hand, forces, and object candidates are expressed in the same camera coordinate system. Keeping rotation fixed, we minimize E_{phys} in Eq. 10 along the viewing ray within ± 25 cm of the current distance at 2 mm intervals. The hand estimate and forces remain fixed throughout refinement. We skip physics correction for inputs containing zero-norm force vectors.

Prior contact with depth and rotation correction. We extend the depth-only objective in Eq. 4 to jointly optimize depth and rotation, adding a quadratic penalty on the rotation update. We optimize ray distance s and rotation increment $\boldsymbol{\omega}$, with $R = \exp([\boldsymbol{\omega}]_{\times})R_0$ and the object's bounding-box center constrained to the current viewing ray. For contact-to-surface distances $d_i(s, \boldsymbol{\omega})$, we minimize

$$E(s, \boldsymbol{\omega}) = \frac{1}{N\nu^2}\sum_{i=1}^{N} \max(d_i(s, \boldsymbol{\omega}) - \delta, 0)^2 + \frac{(s - s_0)^2}{2\sigma_s^2} + \frac{\|\boldsymbol{\omega}\|^2}{2\sigma_R^2}, \tag{12}$$

where N is the number of transferred contact vertices and s_0 is the ray distance at the start of the correction. The contact vertices remain fixed during each correction, while both depth and rotation are updated. We set $\delta = 8$ mm, $\nu = 5$ mm, $\sigma_s = 5$ cm, and $\sigma_R = 10^{\circ}$. Each correction is limited to a 25 cm distance change and a 30° rotation change. We use at most 20 Gauss–Newton steps with backtracking. Initialization first performs the same ± 25 cm, 2 mm contact-distance search as the depth-only variant, then applies the joint correction. Subsequent corrections start from the current MegaPose output.

Table 9: Full results for interaction evidence at proposal and refinement using HIROP-f on HOT3D-Aria. (a) Evidence that ranks the coarse rotations, scored on rotation alone with the ground-truth translation substituted. (b) Evidence that corrects the pose at initialization and again between refiner iterations; all rows share the same proposals and select by MegaPose score. Best values in (b) are bold.

	All (4,402)			Highly occluded (454)		
	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$	$\mathrm{AR}_{\mathrm{ADD}}$	ADD-0.1d	$\mathrm{AR}_{\mathrm{BOP}}$
(a) Proposal						
Visual	60.8	37.1	37.1	54.3	27.5	29.9
Visual + prior grasp (ours)	66.4	38.1	36.1	58.5	33.9	32.3
Visual + prior grasp & contact	64.2	36.1	34.1	56.9	31.5	31.6
Visual + predicted plausibility	66.5	40.0	39.4	60.5	30.0	32.8
(b) Refinement						
Prior position $\rightarrow$ depth	76.0	56.1	62.8	64.8	39.4	51.4
Predicted contact $\rightarrow$ depth	77.4	60.0	65.3	64.2	41.6	52.7
Predicted plausibility $\rightarrow$ depth	62.4	48.9	52.6	58.6	40.5	48.6
Prior contact $\rightarrow$ depth (ours)	**78.2**	**61.9**	**65.9**	**66.0**	**45.2**	**53.4**
Prior contact $\rightarrow$ depth, rotation	77.8	61.5	65.3	64.0	42.7	51.1