INTERGUIDE: GEOMETRIC GUIDANCE FOR VIDEO-BASED LOCO-MANIPULATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling imitable data is critical for learning general-purpose humanoid loco-manipulation, and in-the-wild human–object interaction (HOI) videos offer an economical and virtually unbounded source. However, motion reconstructed from such videos suffers from depth ambiguity, occlusion, temporal jitter, and often lacks detailed hand motion, making the resulting reconstructions difficult to imitate reliably via reinforcement learning (RL). We present **InterGuide**, a framework that leverages geometry-derived interaction guidance to refine HOI reconstruction, and formulate hand-specific rewards to facilitate imitation learning of fine-grained loco-manipulation activities. This guidance takes two complementary forms: pre-sampled coarse gripper poses produced by DenseGripper for grasp interactions, and sparse surface contact anchors for non-grasp interactions (e.g., pushing, supporting). It is integrated across three stages. First, a template-free reconstruction pipeline associates these geometric hypotheses with multi-frame video evidence to refine hand–object contact, depth ordering, and relative scale. Second, interaction-preserving retargeting maps heterogeneous human shapes onto a unified humanoid, refining the palm with the selected gripper pose while aligning gravity and ground support. Third, phase-conditioned grasp rewards use the coarse gripper poses to guide approach, acquisition, and release, while relaxing per-finger tracking to permit contact-adaptive finger motion; non-grasp interactions instead follow the refined contact anchors. Across diverse sequences spanning varied object categories, **InterGuide** reduces reconstructed contact error and delivers substantial gains in hand-motion imitation success rate and grasp fidelity over competitive baselines.

1 INTRODUCTION

Learning from human demonstrations (Xu et al., 2025) is a promising route to humanoid interaction skills, offering natural motion and strong task performance. Yet obtaining high-quality HOI demonstrations at scale remains a major bottleneck. Existing sources each present trade-offs: motion capture is precise but costly, studio-bound, and requires retargeting; teleoperation yields embodiment-ready data but is labor-intensive, hardware-dependent, and limited by robot thermal constraints and imperfect human–robot synchronization; and video generation models (Xie et al., 2026) produce synthetic alternatives but suffer from non-physical artifacts, object-appearance drift over long horizons, and motion confined to the training distribution.

In-the-wild monocular videos—collected from the internet or a single handheld camera—circumvent these limitations. They are economical, virtually unbounded, and inherently physics-consistent, mirroring how humans themselves acquire manipulation skills by observing others (Stefan et al., 2005). However, 3D reconstruction from monocular video is degraded by depth ambiguity, occlusion, and temporal jitter, and fine-grained hand motion such as grasping is difficult to recover reliably from a single viewpoint. Existing imitation frameworks either bypass these ambiguities (Xu et al., 2025) or resort to RL fine-tuning with prohibitive training cost with the imperfect references (Zhao et al., 2026; Huang et al., 2026). Converting physically implausible visual estimates into scalable, physics-valid imitation data therefore remains an open problem.

To address this challenge, we propose **InterGuide**, a framework that converts imperfect monocular HOI reconstructions into imitation signals for humanoid whole-body loco-manipulation (see

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.

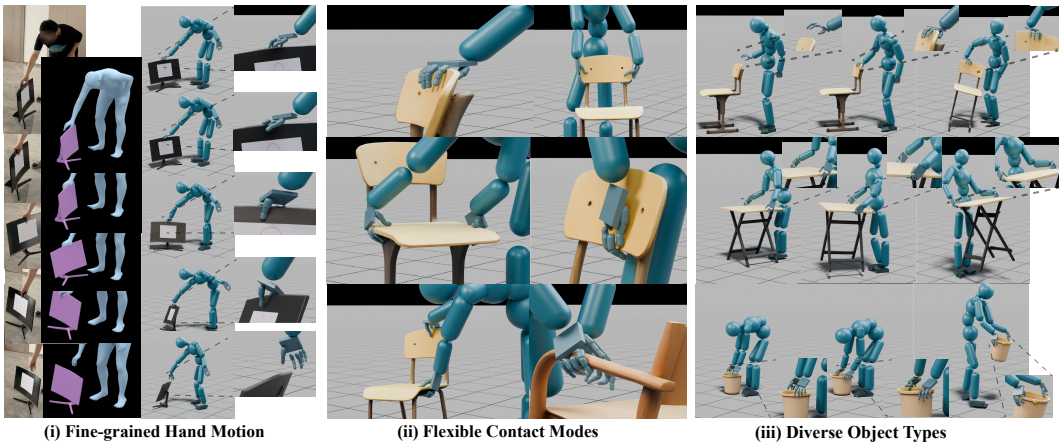

Figure 1: **InterGuide** converts monocular HOI reconstructions into physics-based loco-manipulation motions, enabling (i) fine-grained hand motion, (ii) varied hand–object contact configurations, and (iii) interaction with diverse object types.

Figure 2). Our central idea is to use geometry-derived interaction guidance—pre-sampled simplified grasp pose - gripper poses and sparse surface contact anchors—to (i) refine monocular HOI reconstruction and (ii) formulate effective hand-specific rewards for fine-grained loco-manipulation imitation learning. For initialization, we combine SAM3D (Chen et al., 2026) and FoundationPose (Wen et al., 2024) for object geometry and pose estimation with GVHMR (Shen et al., 2026) and Fast SAM3D Body (Yang et al., 2026) for human motion and hand articulation, all wrapped in one enhanced CARI4D (Xie et al., 2025b) optimization pipeline. Because the human and object are estimated independently, occlusion and depth ambiguity often produce contact disagreement, with hands penetrating or hovering over implausible object regions—precisely the errors our geometric guidance is designed to correct.

We construct two complementary forms of guidance directly from object geometry. For grasp interactions, DenseGripper, inspired by LightningGrasp (Yin & Abbeel, 2025), samples coarse grasp hypotheses—each specifying a grasp center, gripper orientation, and opening width—by searching for compatible surface-normal pairs and testing geometric clearance. Unlike learned grasp detectors (Fang et al., 2023) that commit to sparse predictions, DenseGripper retains dense candidate configurations, enabling downstream selection to remain consistent with observed hand motion. For non-grasp interactions such as pushing or supporting, we sample sparse surface contact anchors together with their accessible approach directions. Given the annotated interaction type, multi-frame hand-keypoint evidence selects the relevant hypotheses and filters outlier matches. The selected anchors then constrain the reconstruction's contact targets to image-supported surface regions. Two additional refinements further improve human–object consistency: part-level depth-order optimization to resolve inter-occlusion ambiguity, and relative-scale adaptation to reduce interpenetration. These geometric hypotheses guide contact correction but do not guarantee a dynamically stable interaction; the downstream policy learns to execute the interaction under simulated dynamics.

To accommodate diverse human body shapes, interaction-preserving retargeting maps the refined motion onto a standardized humanoid. It aligns gravity and ground support while preserving object-local contact relationships for both grasp and non-grasp interactions. For grasping, the selected DenseGripper hypothesis further guides hand-center placement and, where four-finger-side orientation annotations are available, palm-orientation alignment. Even after retargeting, residual hand–object errors leave the policy with a difficult whole-body exploration problem. We therefore convert the selected DenseGripper poses into phase-conditioned rewards for approach, grasp acquisition, and release, while relaxing per-finger tracking to allow contact-adaptive finger motion. These rewards jointly enforce grasp proximity and thumb–finger opposition, complemented by palm-direction and four-finger-side orientation guidance; a geometry-derived termination further prunes episodes with persistent grasp violations. Non-grasp interactions instead follow the refined contact anchors, without imposing an explicit grasp pose guidance. Together, these mechanisms

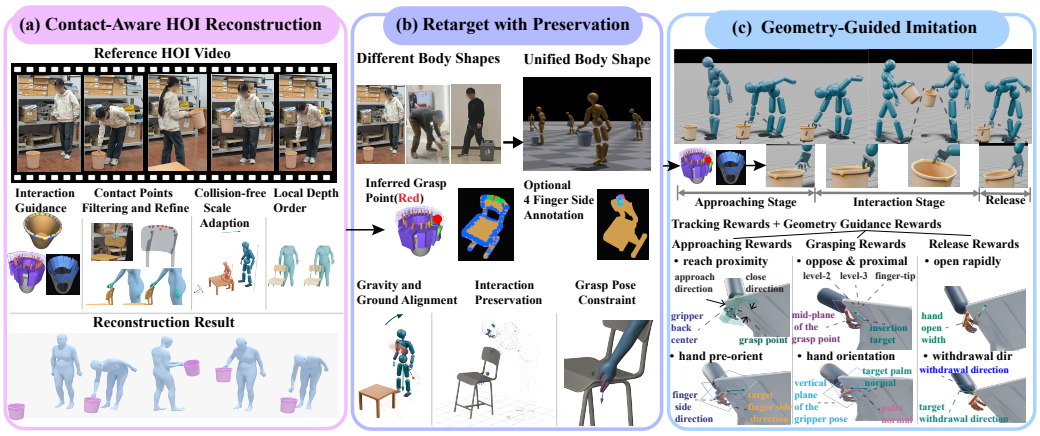

Figure 2: **Overview of `InterGuide`.** (a) Contact-aware reconstruction uses DenseGripper's coarse grasp hypotheses and sparse surface anchors to refine hand–object contact in monocular HOI estimates with relative scale, and depth ordering refinements. (b) Interaction-preserving retargeting transfers the refined motion to a standardized humanoid, aligning gravity and ground support while preserving relative contact geometry and fitting grasp position and palm orientation when finger side annotion is given. (c) Geometry-guided imitation combines tracking rewards with phase-conditioned rewards for approach, grasp, and release. For grasp interactions, coarse gripper poses guide hand exploration while relaxed local hand tracking to adapt detailed finger motion.

reduce the burden of discovering valid interactions through unguided exploration, while leaving detailed contact execution to the physics-based policy.

In summary, our contributions are:

- Contact-aware monocular HOI reconstruction. A single-camera, marker-free pipeline that combines DenseGripper's coarse grasp hypotheses with sparse surface anchors to identify video-supported contact regions, complemented by depth-order constraints and collision-aware relative-scale adaptation.
- Interaction-preserving retargeting. A retarget method that integrates gravity and ground alignment with gripper-guided hand pose fitting, transferring reconstructed interactions to a unified humanoid while preserving object-local contact geometry.
- Geometry-guided imitation learning. A phase-conditioned reward design driven by DenseGripper poses —combining positional and orientation terms with a geometry-derived termination—improves grasp execution from imperfect motion references while allowing flexible finger configurations.

2 Related Work

Monocular 4D HOI reconstruction. VisTracker (Xie et al., 2023) tracks interactions with known object templates, whereas InterTrack (Xie et al., 2025a) jointly estimates object shape and motion without predefined templates. CARI4D (Xie et al., 2025b) combines foundation-model predictions, pose-hypothesis selection, and contact refinement for category-agnostic reconstruction at metric scale. HA-HOI (Zhao et al., 2026) anchors object reconstruction to human motion and projects the recovered interaction into physics-based simulation. RePHO (Huang et al., 2026) refines kinematic estimates through RL, while HSImul3R (Cao et al., 2026) couples human–scene reconstruction with simulation feedback. Building on CARI4D, we further refine hand–object contact using video-supported grasp hypotheses and sparse surface anchors, complemented by depth-order constraints and relative-scale adaptation.

Physics-based imitation and retargeting. PhysHOI (Wang et al., 2023) and SkillMimic (Wang et al., 2025) learn interaction skills from motion demonstrations. InterMimic (Xu et al., 2025) ex-

tends whole-body imitation to diverse object interaction references through interaction-aware tracking, retargeting, and policy distillation. SkillMimic-V2 (Yu et al., 2025) handles sparse and noisy demonstrations via trajectory augmentation and adaptive sampling, while SPIDER (Pan et al., 2025) uses physics-based sampling to retarget human motion across embodiments. We build on InterMimic's imitation framework: our retargeting refinement preserves object-local and hand contacts, and phase-conditioned proximity, opposition, and orientation rewards help the policy accommodate residual reference errors during imitation learning..

Learning from visual demonstrations. VideoMimic (Allshire et al., 2025) learns contextual humanoid control from reconstructed human motion and environments. HumanX (Wang et al., 2026) combines video-derived interaction-data synthesis with unified imitation learning. GRAIL (Xie et al., 2026) generates controllable loco-manipulation demonstration videos from 3D assets and video priors, conditioning reconstruction on known scene geometry, scale, and camera parameters. Huang et al. (2026) combines adaptive sampling with forward and backward policy rollouts to refine noisy kinematic references but offers no guarantee on fine-grained hand behavior. In contrast to previoues works, we operate on in-the-wild monocular videos—where scale and depth ambiguity demand additional refinement—and target fine-grained five-finger loco-manipulation imitation.

3 THE **INTERGUIDE** FRAMEWORK

Given a monocular video of a human interacting with a rigid object, **InterGuide** learns a physics-based policy that reproduces the demonstrated interaction from reconstruction. The pipeline comprises three stages (Figure 2): contact-aware reconstruction refines the estimated human–object motion (Section 3.1); interaction-preserving retargeting transfers it to a unified humanoid (Section 3.2); and geometry-guided imitation learns to execute it in simulation (Section 3.3). The stages share a set of interaction hypotheses derived from object geometry: coarse gripper poses for grasping and accessible surface anchors for non-grasp contact. These hypotheses connect what the object geometry permits with the interaction observed in the video, without prescribing detailed finger motion. We assume annotated contact intervals, with optional finger-gripper-side annotations for signed orientation guidance (Appendix C).

3.1 CONTACT-AWARE HOI RECONSTRUCTION

Monocular estimates can explain the image while placing the hand on the wrong object region or misaligning human–object depth and scale. We address these ambiguities by associating geometric interaction hypotheses with multi-frame video evidence, then using the resulting contact regions to constrain reconstruction.

Motion initialization. Building on CARI4D (Xie et al., 2025b), we initialize object geometry with SAM3D (Chen et al., 2026), scale with monocular depth (Piccinelli et al., 2024), and object poses with FoundationPose (Wen et al., 2024). GVHMR (Shen et al., 2026) supplies the human motion. We preserve its root and lower-body trajectory while refining the torso, arms, and object translation, so contact correction does not overwrite the estimated locomotion. Body-object refinement operates in SMPL-H; the result is subsequently transferred to SMPL-X and combined with the hand articulation estimated by Fast SAM3D Body (Yang et al., 2026).

Geometry-derived interaction hypotheses. We use the reconstructed object geometry to infer valid interaction positions. For grasping, **DenseGripper** constructs a bank of coarse gripper poses around object geometry (Figure 3a). It first generates opposing-surface pairs, filters them by approach accessibility and local shape, and suppresses redundant grasp centers by proximity. At each remaining center, it searches closing and approach directions using a U-shaped virtual gripper, retaining only centers that admit a pose passing a proxy collision test. Each hypothesis specifies a grasp center, orientation, and opening; a single grasp point may yield multiple gripper poses when several approaches are feasible (see the green poses in the chair case). An opposition mid-plane, derived from the local geometry at each grasp point, later guides fingers opposition. For non-grasp interactions such as pushing or supporting, we instead construct sparse surface anchors (Figure 3b). A surface sample is retained only if a finite-size proxy can reach it through a clear straight corridor, which prevents interior points from being selected on non-watertight meshes. Normal-aware sparsification then removes redundant samples while preserving distinct object faces. Unlike grasp hypotheses,

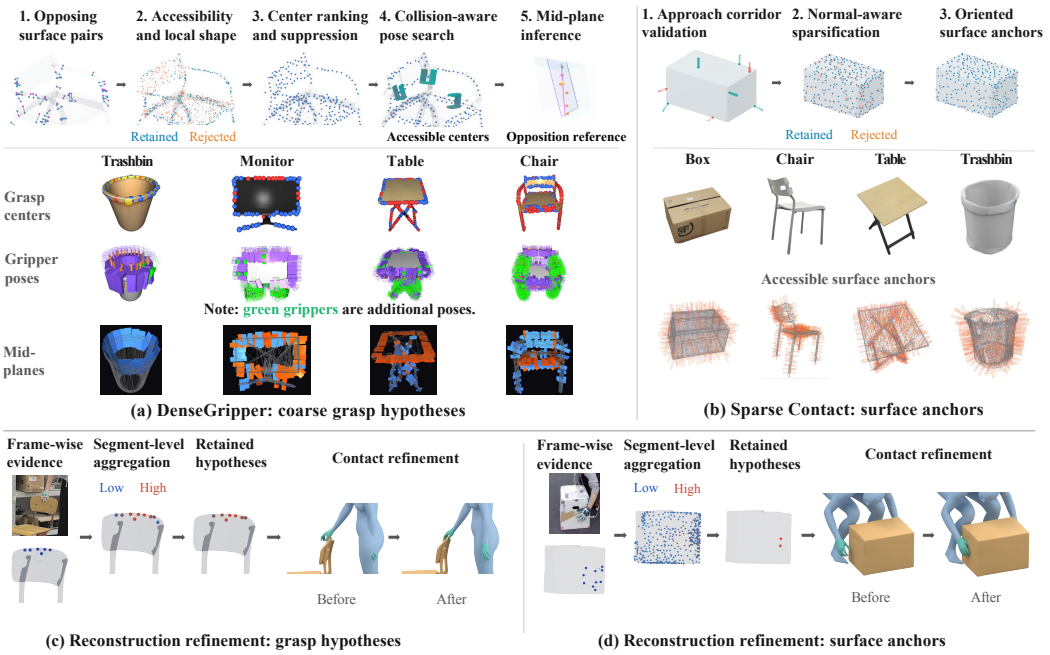

Figure 3: **Geometry-derived interaction hypotheses.** (a) Pipeline and example results of DenseGripper; local geometry further defines an opposition mid-plane. (b) Pipeline and example results of Sparse Contact. (c) Reconstruction refinement with grasp-point hypothesis voting and filtering. (d) Reconstruction refinement with sparse contact-point hypothesis voting and filtering.

these anchors specify an accessible contact location without imposing finger opposition or a grasp pose. Both representations describe geometric interaction accessibility; construction details appear in Appendices A.2 and A.3.

Video-conditioned contact refinement. Object geometry alone does not identify the demonstrated interaction. Within each annotated hand-contact interval, we associate confident image keypoints (Xu et al., 2022) with the rendered visible object surface and accumulate support for nearby hypotheses across frames (Figure 3c,d). Filtering hypotheses by relative frequency to the strongest candidate yields the admissible contact region: surface neighborhoods around retained grasp centers for grasping, and the retained anchors for non-grasp contact. This aggregation reduces sensitivity to individual occluded frames while allowing multiple geometric explanations.

Let $\mathcal{C}$ contain hand-contact intervals (C, h), where C uniquely identifies an interval for hand $h \in \{\text{left hand}, \text{right hand}\}$, and let Γ_C be its admissible region in object coordinates. We replace CARI4D's unrestricted nearest-object contact term with distance to this region:

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{base}}^{-\text{cont}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{stat}}\mathcal{L}_{\text{stat}}, \quad \mathcal{L}_{\text{int}} = \sum_{(C,h)\in\mathcal{C}} \sum_{t\in C} d^2\left(x_{t,h}, \Theta_{\mathcal{O}}^{t}\Gamma_C\right). \quad (1)$$

Here $x_{t,h}$ is the SMPLH hand joint, $\Theta_{\mathcal{O}}^{t}$ maps the canonical object to frame t, and d is point-to-set distance. The base objective retains CARI4D's image, silhouette, penetration, temporal, and initialization terms. This constrains hand placement to video-supported object regions.

Resolving scale, depth ambiguity, and object stableness. As illustrated in Figure 2a, a bounded scale search using sampled core-body collision checks adjusts the human scale and translation relative to the fixed object prior to optimization; scale is then held fixed during optimization. We complement contact refinement with segmentation-guided ordinal depth constraints, $\mathcal{L}_{\text{ord}}$, which penalize local human–object depth ordering inconsistent with observed occlusions. For the quasi-static interactions considered here, $\mathcal{L}_{\text{stat}}$ and a subsequent static-segment projection suppress object translation drift in annotated rest intervals. The refined human–object trajectory and a representative grasp point - selected from clip-level evidence by smallest median distance—are then passed to retargeting. Appendix A.1 details the association, loss terms, and optimization schedule.

3.2 Interaction-Preserving Motion Retargeting

Mapping different reconstructed body shapes onto a single fixed humanoid morphology can break contact even when the source motion is well aligned; correcting object jitter or ground penetration creates a similar problem. We therefore preserve object-local interaction geometry while adapting the motion (Figure 2b).

Shared alignment and contact preservation. We first apply a common gravity and ground alignment to the human and object, then smooth object motion and regularize its resting pose towards ground support. To preserve contact through these corrections, we express hand targets in the original object-local frame and map them through the corrected object transform. We then fit sparse SMPL-X keyframes under the target morphology to these transported targets, regularizing deviations from the reference motion. Independent contact masks preserve the inactive arm while the other hand interacts. This construction applies to both grasp and non-grasp contact; its formulation is given in Appendix A.4.

Grasp-pose optimization and optional finger-side annotation. Preserving hand location alone does not determine an appropriate grasp orientation. For grasp clips, in addition to aligning the palm-region center with the selected grasp point, we align the finger direction and the inward palm normal with the gripper-derived axes when four-finger-side orientation annotations are available. The annotated four-finger side resolves the ambiguity between the two opposite sides of the gripper. For grasps not clearly parallel to the gripper's back plane—such as the trashbin pick-up in Figure 2 or the chair-handle and chair-back grasps in Figure 1(ii)—we do not annotate the finger side. Smooth, object-attached target transitions introduce these corrections before contact and blend them into the surrounding motion.

Finally, interpolation of the fitted keyframes yields the humanoid reference $\mathcal{T}_{\mathcal{R}}$. These constraints refine hand placement and orientation; the physics-based policy subsequently adapts the estimated body articulation. Appendix A.4 gives the fitting objective and transition schedules.

3.3 Geometry-Guided Imitation Learning

Although reconstruction and retargeting improve hand–object alignment, the recovered finger motion remains an unreliable target for contact execution. We therefore use the coarse gripper poses to formulate positional and orientation rewards across the approach, grasp, and release phases. The geometry components underlying these rewards are illustrated in Figure 2c. Combined with InterMimic's tracking rewards (Xu et al., 2025), this guidance trains a physics-based policy to imitate $\mathcal{T}_{\mathcal{R}}$ while allowing articulated fingers to adapt flexibly to simulated contact dynamics.

Positional guidance. Before grasping, an approach pre-open reward encourages the hand to reach an opening matched to the gripper width. During early approach, we align the back centers of the hand proxy and the gripper without prescribing individual finger joint angles. Near contact or during contact, guidance shifts from palm-center alignment to balanced thumb–finger proximity. To give both sides of a grasp equal influence, we balance the thumb distance against the mean distance of the index, middle, and ring fingers. Omitting time and hand indices, let $p_{i,\ell}$ be finger i's rigid-body origin at joint level $\ell \in \{2,3\}$ and c the inferred grasp center, both in world coordinates. We define

$$D = \min_{\ell\in\{2,3\}} \frac{1}{2}\left(\|p_{\mathrm{th},\ell} - c\|_2 + \frac{1}{3}\sum_{i\in\mathcal{F}_3}\|p_{i,\ell} - c\|_2\right), \qquad r^{\mathrm{prox}} = e^{-\lambda_f D}, \tag{2}$$

where $\mathcal{F}_3$ contains the index, middle, and ring fingers. The minimum allows either tested joint level to establish proximity; palm-center guidance uses the same exponential form with a single distance.

Proximity alone can reward one-sided touching. During grasp, we therefore combine finger contact with an opposition reward defined by the geometry-derived mid-plane. Let b_i indicate whether the thumb and finger i occupy opposite sides at either tested joint level (2 and 3), and let $\mathcal{F}_4$ contain all four non-thumb fingers. For clips with a four-finger-side annotation,

$$Q = \tfrac{1}{4}\sum_{i\in\mathcal{F}_4} b_i, \qquad r^{\mathrm{opp}} = e^{-\lambda_d(1-Q)}\,\frac{1+H}{2}, \tag{3}$$

where Q is the fraction of opposing fingers, and depth factor $H = 1$ if at least one opposing thumb–finger pair has both physical fingertips sufficiently past the selected grasp center along the insertion direction, and $H = 0$ otherwise. This rewards multi-finger opposition together with sufficient insertion, rather than mere proximity. Without the annotation, we reward any opposing pair and omit the depth factor (Appendix A.5). Once contact ends, positional guidance instead encourages thumb–finger opening.

Orientation guidance. Proximity and opposition do not determine palm facing or finger-side assignment. We therefore introduce two complementary orientation rewards active during approach and grasping. The palm-direction reward, r^{palm}, aligns the inward palm normal with the incoming gripper direction projected onto the mid-plane. The four-finger-side reward, r^{side}, aligns a wrist-fixed axis representing the four-finger side with the signed gripper closing direction specified by the annotated rail. Both rewards use the same tolerance-aware form for pose flexibility:

$$r^{j} = \exp\left[-\lambda_j \alpha_t^j \frac{[\cos\theta_{to} - v_j^\top v_j^\star]_+}{1 + \cos\theta_{to}}\right], \qquad j \in \{\text{palm}, \text{side}\}, \tag{4}$$

where $v_j, v_j^\star$ are unit current and target directions, θ_{to} is the angular tolerance, λ_j sets the penalty scale, and α_t^j controls phase activation. These terms encourage pre-orientation and grasp maintenance without prescribing finger joint angles. When the finger-side annotation is unavailable, both signed factors default to one. An additional unsigned factor encourages the palm-normal direction to lie within the gripper's longitudinal plane, with or without an annotation (Appendix A.5). This plane corresponds to the "vertical plane" in Figure 2c and rotates with the gripper rather than following world gravity. During release, we additionally reward object-local wrist withdrawal along the outgoing gripper direction.

Coupling guidance with imitation. The phase-specific hand-guidance factors combine multiplicatively with the tracking reward, with inactive-phase factors set to one (e.g., $r_{t,h}^{app} = 1$ during grasp).

$$r_t = \tilde{r}_t^{\text{track}} \prod_{h \in \{L,R\}} r_{t,h}^{\text{app}}\, r_{t,h}^{\text{grasp}}\, r_{t,h}^{\text{rel}}. \tag{5}$$

The term $\tilde{r}_t^{\text{track}}$ retains body, object, and non-hand contact supervision. Non-grasp sequences follow the refined contact references without gripper-pose or opposition rewards.

Geometry-derived termination. Body and object tracking rewards can admit prolonged rollouts with poor grasps. We therefore augment tracking-based termination with checks for persistent opposition, grasp-proximity, and palm-direction violations. Appendix A.5 specifies the remaining reward formulas, hyperparameters, phase schedules, and termination conditions.

4 Experiments

We design our experiments to answer five questions. **(Q1)** Does geometry-guided contact refinement effectively correct reconstructed contacts and **(Q2)** how does it affect the following imitation learning? **(Q3)** Does **`InterGuide`** improve fine-grained hand imitation compared with InterMimic and a GRAIL-inspired reward baseline? **(Q4)** How much do individual components—such as the orientation rewards and the geometry-derived termination—contribute? **(Q5)** What is the preferred training regime for **`InterGuide`**, single or multi-sequence learning?

Data and baselines. Our evaluation set comprises six non-grasp box sequences from BEHAVE (Bhatnagar et al., 2022) and 25 hand-grasp self-captured sequences covering chairs, tables, monitors, and trashbins, spanning a diverse range of interaction modes. To assess reconstruction quality, we compare two variants of our enhanced CARI4D, with and without contact refinement. For imitation in simulation, we adopt InterMimic as our primary baseline and additionally construct **RGrailHand**, which augments InterMimic's tracking rewards with GRAIL-inspired hand rewards (Xie et al., 2026) that encourage finger contact count, finger opposition, and grasp proximity. All baselines and **`InterGuide`** share InterMimic's policy architecture, rather than GRAIL's. This controls for differences in policy architecture..

Metrics. For reconstruction, *first-contact error* is the distance between the reconstructed hand position at contact onset and its annotated ground truth first-grasp point, and *contact-interval error* averages this distance over the entire contact interval. For imitation, *grasp success* requires sustained

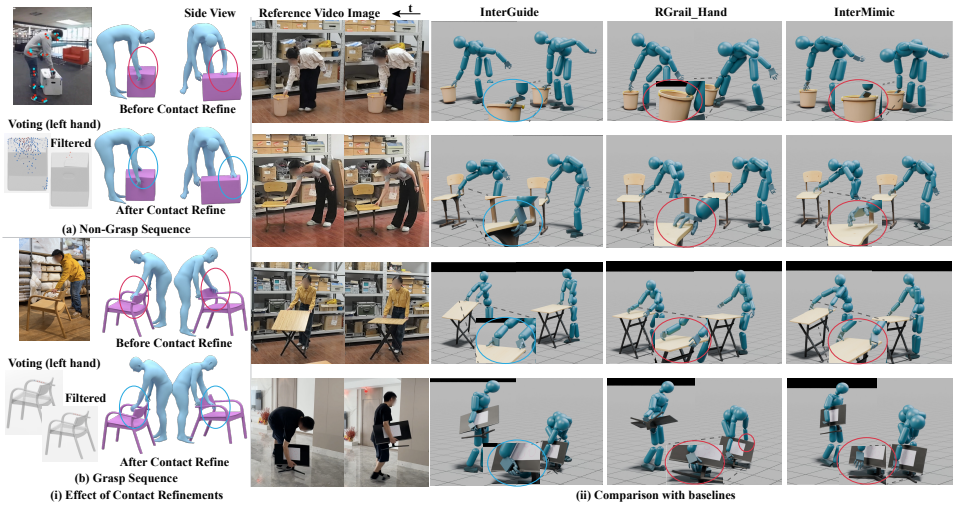

Figure 4: **Comparison with baselines.** The geometry-derived contact refinement and rewards reduce reconstruction errors and attain more grasp success across different interactions and objects.

Table 1: **Effect of contact-aware reconstruction refinement.** First-contact error measures the Euclidean distance between each hand's position at contact onset and its ground-truth first-grasp point. Contact-interval error averages this distance over the corresponding frames. "w/o ref." and "w/ ref." denote reconstruction without and with refinement, respectively.

Metric (cm) ↓	Chair		Table		Monitor		All	
	w/o ref.	w/ ref.	w/o ref.	w/ ref.	w/o ref.	w/ ref.	w/o ref.	w/ ref.
First-contact error	10.50	**8.14**	12.71	**8.27**	5.92	**5.70**	10.52	**7.81**
Contact-interval error	9.31	**6.36**	9.19	**6.85**	9.57	**8.23**	9.37	**7.04**

finger opposition and grasp proximity, together with palm-direction and four-finger-side alignment when annotated, and completion of the corresponding contact segment. *Rollout success* measures completion of the corresponding reference segment, irrespective of grasp quality. To accommodate sequences with different numbers of contact segments, we score each sequence by its fraction of successful segments and report n/N, where N is the number of sequences and n sums these fractional contributions.

Grasp proximity (GProxi) reports the balanced thumb–finger distance D from Equation 2, averaged over hand-contact frames labeled as grasps in the reference. *Contact preservation* measures the agreement between reference and simulated hand-contact states across reference-contact frames. Finally, *body* and *object tracking errors* measure the root drift of the body and the object, respectively, with respect to the reference motion.

Q1:Reconstruction refinement improves grasp alignment. Table 1 shows consistently lower errors with refinement across all object categories: first-contact error decreases from 10.52 to $7.81\,\mathrm{cm}$ (25.8%), while contact-interval error decreases from 9.37 to $7.04\,\mathrm{cm}$ (24.9%), indicating that refinement improves both the initial hand placement and its sustained alignment with the intended grasp region. Figure 4(i) visualizes representative improvements.

Q2:Refined references improve non-grasp imitation. On box interactions (both use multi-sequence learning), refining the reference's contact lifts InterMimic's rollout success from 83.3% to 100%, and reduces body and object tracking errors from 9.48, 19.87 to 7.22, $11.36\,\mathrm{cm}$.

Q3:Geometric guidance improves grasp execution. Under the same single-sequence imitation-learning setup used by all methods, **`InterGuide`** attains the highest grasp success and lowest GProxi across all 4 categories– 87.5% (chairs), 58.3% (tables), 100% (monitors), and 87.5% (trashbins). Rollout success tells a different story: **`InterGuide`** trails the best baseline on chairs and tables (87.5% vs. 90.0%; 58.3% vs. 83.3%), and on trashbins all methods reach 100% rollout suc-

Table 2: **Imitation performance by object type and Ablation Study**. 'B' means multi-sequence learning while the others are single sequence learning. For Grasp and rollout success, we report the percentage with n/N in parentheses, where N is the number of evaluated sequences and n is the sum of their fractions of successful contact segments. GProxi is Grasp Proximity. "–" denotes a not-applicable measurement. ROrient denotes orientational geometry guidance rewards while GTerm is the geometry-derived termination.

Object type	Method	Success (%,n/N)		Contact preservation (%)	GProxi (cm)	Body track (cm)	Object track (cm)
		Grasp	Rollout				
Box	No C-ref.(B)	–	83.3 (5/6)	99.63	–	9.48	19.87
	C-ref.(B)	–	**100.0 (6/6)**	99.85	–	**7.22**	**11.36**
Chair	InterMimic	35.0(3.50/10)	**90.0(9.00/10)**	95.44	11.81	**8.98**	**9.78**
	RGrailHand	32.5(3.25/10)	**90.0(9.00/10)**	94.19	11.72	9.64	10.56
	InterGuide	**87.5(8.75/10)**	87.5(8.75/10)	**98.14**	**6.44**	11.27	13.09
Table	InterMimic	33.3 (2.0/6)	**83.3 (5.0/6)**	**95.79**	12.29	9.50	10.10
	RGrailHand	25.0 (1.5/6)	**83.3 (5.0/6)**	95.42	12.39	**9.41**	**9.91**
	InterGuide	**58.3 (3.5/6)**	58.3 (3.50/6)	92.06	**7.44**	11.01	11.38
Monitor	InterMimic	20 (1/5)	40 (2/5)	86.45	9.94	11.47	15.47
	RGrailHand	40 (2/5)	60 (3/5)	97.34	6.49	11.04	15.48
	InterGuide	**100 (5/5)**	**100 (5/5)**	**99.27**	**5.67**	**10.60**	**11.93**
Trashbin	InterMimic	25.0 (1.0/4)	100 (4/4)	88.74	12.8	11.79	**11.66**
	RGrailHand	62.5 (2.5/4)	100 (4/4)	94.95	7.62	12.82	12.60
	InterGuide	**87.5 (3.5/4)**	100 (4/4)	**98.95**	**6.92**	**11.74**	12.02
Chair-6 for Ablation	InterMimic	8.33 (0.5/6)	83.3 (5.0/6)	95.74	13.15	**9.81**	**11.62**
	w/o ROrient	16.7 (1.0/6)	16.7 (1.0/6)	88.16	13.28	14.23	13.90
	w/o GTerm	16.7 (1.0/6)	83.3 (5.0/6)	89.92	16.75	12.89	11.75
	InterGuide	**91.7 (5.5/6)**	**91.7 (5.5/6)**	**98.07**	**6.79**	11.39	13.90

cess while grasp success ranges from 25.0% to 87.5%. These contrasts show that rollout success alone can mask failures to establish the intended grasp, which our geometry-derived rewards and termination are designed to prevent. The lower success rate on tables is primarily due to walking collisions before the second contact, which we find can be alleviated by multi-sequence learning (**Q5**). Figure 4(ii) further demonstrates the improved grasp fidelity from **`InterGuide`**.

Q4:Component ablations. On six representative chair sequences, we ablate orientation guidance (ROrient) and geometry-derived termination (GTerm); results are reported in Table 2. Removing either component reduces grasp success from 91.7% to 16.7%, indicating that both contribute substantially on this subset. Without GTerm, rollout success remains 83.3% despite low grasp success, highlighting the distinction between rollout completion and grasp quality. **Q5** is addressed in Appendix B.

5 Conclusion

We presented **`InterGuide`**, a framework for learning humanoid loco-manipulation from imperfect monocular HOI reconstructions. DenseGripper hypotheses together with sparse surface anchors provide complementary geometric guidance for grasp and non-grasp interactions: this guidance refines reconstructed contact, is preserved through retargeting, and shapes phase-conditioned hand rewards, while still letting the policy adapt finger-level motion in simulation. Experiments demonstrate reduced reconstruction contact errors, improved non-grasp tracking, and consistently higher grasp success across the evaluated object categories. Together, these results establish geometry-derived guidance as an effective means of turning imperfect video reconstructions into a usable source of supervision for physics-based imitation. Reducing annotation requirements and extending **`InterGuide`** to more dynamic, long-horizon interactions are important directions for future work. Transferring these imitated results to physical robots is also an appealing next step.

AI USE STATEMENT

In this work, we used LLM-based tools to retrieve recent related work, assist in drafting parts of the Method section, and polish the writing. Introduction, Related Works, and Experiments are drafted by the authors. DenseGripper was initially developed by the authors and subsequently refined with the assistance of LLM-based analysis. All other research ideas and methodological components were developed by the authors. The reconstruction refinements, retargeting, and imitation rewards were designed entirely by the authors, while LLM-based coding tools assisted with parts of their code implementation. We have reviewed all AI-assisted content and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies humanoid loco-manipulation in simulation using human–object interaction videos from existing datasets and self-captured recordings. We obtained participants' permission to use and publish their recordings and blurred their faces to reduce the risk of identification. Our simulation results do not establish the safety of deployment on physical robots; real-world use requires additional validation and safeguards.

REPRODUCIBILITY STATEMENT

We describe the reconstruction, retargeting, and imitation-learning framework in Section 3. Appendix A provides algorithmic and implementation details, including geometric hypothesis construction, coordinate transformations, optimization procedures, reward components, and termination conditions. Appendix C documents the manual annotations used in the pipeline. Section 4 and Appendix B present the experimental results and describe the evaluation metrics and aggregation procedures, including the comparison between single- and multi-sequence training. Code and data will be made publicly available to facilitate reproducibility and further research.

REFERENCES

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yukang Cao, Haozhe Xie, Fangzhou Hong, Long Zhuo, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Hsimul3r: Physics-in-the-loop reconstruction of simulation-ready human-scene interactions, 2026.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026.

Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

Dingbang Huang, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Recovering physically plausible human-object interactions from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7079–7088, 2026.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, et al. Dposer-x: Diffusion model as robust 3d whole-body human pose prior. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9988–9997. IEEE, 2025.

Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. Spider: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116. IEEE, 2024.

Zehong Shen, Jiawei Wang, Huaijin Pi, Yan Xia, Zhi Cen, Zechen Hu, Hujun Bao, Ruizhen Hu, Xiaowei Zhou, and Sida Peng. Gvhmr: Gravity-view coordinates for global human motion recovery from monocular videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Katja Stefan, Leonardo G Cohen, Julie Duque, Riccardo Mazzocchio, Pablo Celnik, Lumy Sawaki, Leslie Ungerleider, and Joseph Classen. Formation of a motor memory by action observation. *Journal of Neuroscience*, 2005.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17540–17549, 2025.

Yinhuai Wang, Qihan Zhao, Yuen Fui Lau, Runyi Yu, Hok Wai Tsui, Qifeng Chen, Jingbo Wang, Jiangmiao Pang, and Ping Tan. Humanx: Toward agile and generalizable humanoid interaction skills from human videos. *arXiv preprint arXiv:2602.02473*, 2026.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17868–17879, 2024.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, et al. Grail: Generating humanoid loco-manipulation from 3d assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4757–4768, 2023.

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. In *2025 International Conference on 3D Vision (3DV)*, 2025a.

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. Cari4d: Category agnostic 4d reconstruction of human-object interaction. *arXiv preprint arXiv:2512.11988*, 2025b.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Timing Yang, Sicheng He, Hongyi Jing, Jiawei Yang, Zhijian Liu, Chuhang Zou, and Yue Wang. Fast sam 3d body: Accelerating sam 3d body for real-time full-body human mesh recovery. *arXiv preprint arXiv:2603.15603*, 2026.

Zhao-Heng Yin and Pieter Abbeel. Lightning grasp: High performance procedural grasp synthesis with contact fields. *arXiv preprint arXiv:2511.07418*, 2025.

Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Yubo Zhao, Yujin Chai, Yunao Dong, Chengfeng Zhao, Zijiao Zeng, Yuan Liu, and Chi-Keung Tang. Real2sim in hoi: Toward physically plausible hoi reconstruction from monocular videos. *arXiv preprint arXiv:2605.14462*, 2026.

APPENDIX

We provide reconstruction, optimization, and imitation learning's details for the geometric guidance introduced in Section 3, followed by the training configuration, evaluation protocol, and additional results for **Q5** (Appendix B), and the annotation procedure (Appendix C).

A ADDITIONAL METHOD DETAILS

A.1 RECONSTRUCTION REFINEMENT

This subsection details Section 3.1 and Figure 3c,d.

Relative-scale search (Figure 2a). Before motion optimization, human scale is searched over $[1, 1.2]$ in increments of 0.02 relative to the fixed object. Human geometry and its translation in camera coordinates are scaled together about the camera origin, preserving image projection. We test 1,000 object-surface samples against a core-body signed-distance field, excluding arms, hands, and feet to avoid penalizing intended contact. The smallest passing scale is retained, or the upper bound if no candidate passes, and the check is repeated after body-shape fitting. The test uses frames up to first hand contact, or reliable object-foreground frames for non-contact sequences. Scale is then fixed during motion optimization. This bounded search reduces core-body penetration without guaranteeing collision-free motion.

Optimized variables. Equation 1 jointly optimizes the object pose and the human SMPL-H body. It refines the torso, arms, and object translation while retaining the GVHMR root, lower-body, neck, and head motion. SMPL-H optimization uses mean hand poses and one joint per hand to correct placement rather than finger articulation. The result is transferred to SMPL-X with global pose and shape retained, after which estimated hand articulation from SAM3D Body (Yang et al., 2026) is restored.

Video-evidence voting for Γ_C. Illustrated in Figure 3, confident image keypoints are matched to projections of the visible object surface. Support is accumulated within each hand-contact interval, with at most one vote per hypothesis per frame. For grasping, a surface observation supports centers within $5\,\mathrm{cm}$, or the nearest center if none falls within this radius. Centers receiving at least 40% of the maximum candidate support are retained; their $8\,\mathrm{cm}$ surface neighborhoods define the admissible contact region. For non-grasp interactions, all 21 hand keypoints contribute evidence. Each observation supports only its nearest surface anchor within $3\,\mathrm{cm}$, and the relative-support threshold is 80% since the sparse contact anchors are denser than grasp points. Both thresholds are relative to the best-supported hypothesis, not to the number of frames. After refinement, the center whose median hand distance over the evidence frames is smallest provides the representative grasp anchor c^{first} for retargeting; the median is used to guard against outliers. After annotation verification (Appendix C), its saved per-segment index also specifies the fixed grasp point target used to evaluate grasp proximity (Appendix B).

Ordinal depth $\mathcal{L}_{\mathrm{ord}}$. At a projected human anchor, the observed segmentation determines which surface should be in front. Inconsistent ordering is penalized by

$$\mathcal{L}_{\mathrm{ord}} = \sum_{t,i} w_{t,i}\, \rho\Big([m_{t,i} - \sigma_{t,i}(z_{\mathcal{O}}^{t,i} - z_{\mathcal{H}}^{t,i})]_+\Big), \tag{6}$$

where $z_{\mathcal{O}}^{t,i}$ and $z_{\mathcal{H}}^{t,i}$ are local object and human depths, and $\sigma_{t,i}$ is $+1$ for human-foreground and -1 for object-foreground evidence. Here, ρ is a Huber penalty applied to the nonnegative margin violation, with transition scale $\delta = 5\,\mathrm{mm}$. The positive-part operator enforces the contact-aware margin $m_{t,i}$; robust penalty ρ and confidence weights $w_{t,i}$ suppress ambiguous masks and depth discontinuities. The loss acts on forearm vertices and hand joints after a warm-up.

Rest-phase stabilization $\mathcal{L}_{\mathrm{stat}}$. In the quasi-static setting considered here, we restrict our scope to objects that undergo little dynamic motion from hands interaction, and treat annotated object no-contact intervals as rest phases. During the first half of refinement, a soft penalty discourages

frame-to-frame translational motion within these intervals. During the latter half optimization, this penalty is replaced by a projection after each optimizer update. Let $\widetilde{\mathbf{p}}_t$ denote the updated object translation and $\mathcal{S}_j$ a rest interval. We enforce

$$\mathbf{p}_t \leftarrow \begin{cases} \widetilde{\mathbf{p}}_{a_j}, & j \in \mathcal{A}, \\ \dfrac{1}{|\mathcal{S}_j|} \displaystyle\sum_{\tau \in \mathcal{S}_j} \widetilde{\mathbf{p}}_\tau, & \text{otherwise}, \end{cases} \qquad t \in \mathcal{S}_j, \tag{7}$$

where $\mathcal{A}$ contains intervals with a subsequent contact-boundary anchor a_j. Here, $\mathcal{A}$ indexes rest intervals followed by contact, and a_j denotes the first frame of that subsequent contact. For these intervals, all object translations are set to the updated translation at frame a_j, preserving positional continuity at contact onset. Intervals without a subsequent contact use their current mean translation instead. Recomputing this shared value after every update allows each interval to shift as a whole while eliminating intra-interval translational motion. Object rotations are not optimized and retain the FoundationPose estimates.

Optimization settings. Reconstruction is configured for 3,000 steps with batches of 64 consecutive frames, using Adam with initial learning rate 10^{-3} and a cosine schedule with 300 warm-up steps. The contact and static-velocity weights (λ_{int}, λ_{stat}) are 200 and 600, respectively; depth ordering (λ_{ord}) uses 200. The inherited 2D-joint, silhouette, temporal second-difference, penetration, and initial-object- translation weights in $\mathcal{L}_{\text{base}}^{-\text{cont}}$ are 0.03, 0.002, 1000, 2, and 100. Depth ordering starts after the first third of optimization. At the midpoint, static-translation projection replaces the soft static-velocity penalty; penetration is activated after 60% of the steps.

A.2 DenseGripper Construction

DenseGripper constructs an object-local bank of coarse grasp hypotheses, comprising grasp centers, collision-checked gripper poses, and opposition mid-planes. Its five construction stages follow Figure 3a: opposing surface-pair proposals, accessibility and local-shape filtering, center ranking and suppression, collision-aware pose search, and mid-plane inference.

Opposing surface pairs. We draw 10,000 area-uniform surface samples, voxel-subsample them at 1 cm, and select at most 3,000 ray origins by farthest-point sampling. Each origin casts an inward-normal ray and five additional rays on a 30° cone. An opposing surface pair is accepted when its separation is 0.5–15 cm, its normal dot product is below -0.5, and the absolute alignment between the pair axis and each surface normal exceeds 0.3. The pair midpoint, unit axis, and separation define a candidate grasp center c, closing direction d, and opening w. Unlike the surface anchors in Figure 3b, a grasp center may lie inside the object: it identifies a region to enclose rather than a surface point to touch.

Accessibility and local shape. Each pair proposal is screened for accessibility using 16 approach directions orthogonal to d. A thin proxy with a 1 cm-square cross section and 6 cm length tests the approach corridor through center rays, a cross-sectional ray grid, and swept-containment checks. Local PCA and boundary-normal consistency further favor rod-like regions and sheet edges or rims over bulky shapes. This stage removes obstructed or geometrically unsuitable proposals before the more detailed gripper-pose search.

Center ranking and suppression. Surviving proposals are ranked by normal opposition, pair-axis alignment, accessibility, local shape, and opening. Centers within 3 cm of a higher-ranked center are suppressed to reduce redundancy. The retained centers remain provisional until a compatible gripper pose passes the subsequent clearance checks.

Collision-aware pose search. For each retained center, local surface normals provide candidate closing axes. We test 16 palm-placement directions per axis using the U-shaped virtual gripper introduced in Section 3.1 and Figure 3a, with 2 cm rail/back thickness, 12 cm depth, and 5 cm height to resemble human hand sizes. A placement is rejected if sampled surface points or mesh vertices intersect its rails or back block. A center survives only if at least one tested pose passes the clearance

checks, yielding accessible grasp centers and their associated gripper poses. Multiple valid poses may share the same center (see the green gripper poses around the handles of chair in Figure 3a).

For each pose, direction a points from the grasp center toward the palm; the fingers extend along $-a$, and the pre-grasp target is displaced by $6\,\mathrm{cm}$ along a. These sampling-based checks approximate geometric clearance rather than exact articulated-hand collision checking.

Mid-plane inference. For each grasp center c, we construct a plane for testing whether the thumb and other fingers lie on opposite sides of the grasp. We first collect near-surface points inside the object and within the grasp regions of accepted gripper poses, then fit a line with direction ℓ to describe the local object geometry. The mid-plane passes through c and spans ℓ and $z = d \times a$, giving the unit normal $n = (z \times \ell)/\|z \times \ell\|$. If the line fit is degenerate, we instead use a plane perpendicular to the gripper closing axis. We also check consistency with the accepted gripper poses: if the fitted normal deviates by more than 45° from every accepted closing axis, we replace the plane with the center plane of a compatible gripper. The resulting plane supplies the opposition reference in Equation 3. Figure 3a shows several examples of grasp points, gripper poses, and mid-planes

Offline construction and use. The geometric bank is constructed offline and remains fixed during policy training. Video evidence selects contact-relevant hypotheses for reconstruction refinement (Figure 3c). At imitation learning runtime, a saved center-level mid-plane is used when its coordinate frame matches and the center admits one to three successful poses; otherwise, the selected gripper's plane is used.

A.3 Sparse Surface Anchors

Sparse surface anchors provide accessible contact targets for non-grasp interactions (Figure 3b). Each anchor specifies a surface location and an accessible approach direction, without prescribing a gripper pose. It follows the following three steps:

Candidate sampling. We sample 50,000 points on the object mesh and favor points on smooth surface patches, where neighboring normals point in similar directions and the points lie away from boundaries and sharp creases. An initial filtering step removes redundant candidates within $9\,\mathrm{mm}$.

Approach corridor validation. For each candidate, we check whether a thin contact proxy can reach the point from outside the object without obstruction to remove interior contact points. We test five possible approach corridors: one extending along the outward surface normal and four tilted by 20° around it. Each corridor has a $1\,\mathrm{cm}$-square cross section, is at least $6\,\mathrm{cm}$ long, and extends beyond the padded object bounding box. Rays through the corridor center and cross section first detect obstructions. Additional intersection and containment tests check whether the object blocks the volume swept by the proxy. For open meshes, we also test access from the opposite side if the preferred side is blocked. A candidate is retained if at least one corridor is clear, and its first feasible direction is stored.

Normal-aware sparsification. Accessible candidates are ranked by approach clearance and surface-patch quality. We then retain higher-ranked points and remove redundant neighbors within $2\,\mathrm{cm}$. In both this step and the initial $9\,\mathrm{mm}$ filtering, a neighbor is removed only if its surface normal differs by at most 60° from that of the retained point (normal dot product at least 0.5). Thus, nearby points on sufficiently different faces are preserved rather than merged. The resulting anchors remain on the original object surface, unlike DenseGripper centers, which may lie inside the object. Figure 3b shows several sparse-contact examples.

A.4 Retargeting

Retargeting (Section 3.2, Figure 2b) aims to preserve the reconstructed human–object interaction relationship while refining contact and hand-pose details.

Gravity and ground correction. Shared gravity and ground corrections are applied to the human and object, with the floor estimated from the first ten frames of human feet's heights. Gravity

alignment corrects the vertical orientation of the reconstructed meshes using the gravity direction predicted by GVHMR, as illustrated in Figure 2b. Object translation and rotation are smoothed using constant-velocity filters, and rest intervals are assigned a fixed pose. These corrections establish the object trajectory used for subsequent human fitting.

Interaction preservation. To preserve contact while correcting the object trajectory, we retain each contacting hand target in object-local coordinates. Let F_t and $\tilde{F}_t$ denote the object transforms before and after correction. An original hand target $x_{t,h}$ is mapped into the original object frame and then transported through the corrected transform:

$$\xi_{t,h} = F_t^{-1}(x_{t,h}), \qquad \tilde{x}_{t,h} = \tilde{F}_t(\xi_{t,h}). \tag{8}$$

Thus, the target follows the corrected object while retaining its object-local contact location. The transforms use the convention

$$F_t(\xi) = s_t \odot (R_t\xi) + p_t, \qquad F_t^{-1}(x) = R_t^\top\big((x - p_t) \oslash s_t\big), \tag{9}$$

where $\odot$ and $\oslash$ denote element-wise multiplication and division. If the mesh is centered by its vertex mean c_{mesh}, both vertices and anchors are shifted by $-c_{\mathrm{mesh}}$, and the translation is compensated as $\bar{p}_t = p_t + s_t \odot (R_t c_{\mathrm{mesh}})$. This change of coordinate origin preserves the world-space geometry.

Body and palm fitting. We fit body pose under a 16-coefficient fixed lower-body target morphology (Lu et al., 2025), keeping the aligned root motion and articulated finger poses fixed. Keyframes include the endpoints, every third frame, and contact transition boundaries. Independent contact masks condition each hand's fitting targets on its own interaction schedule. Ordinary targets constrain the wrist, whereas grasp-center targets (c^{first}) constrain the mean of designated palm vertices; regularization limits departures from the reference pose.

To avoid abruptly switching the palm target at contact onset, we gradually introduce the first-grasp positional correction over the preceding window of up to ten frames. The correction is expressed in object-local coordinates and transported by the object transform at each frame. It is added to the original palm trajectory with weight $\eta(u) = 6u^5 - 15u^4 + 10u^3$, where u increases from zero at the window start to one at contact onset. Thus, the original approach motion is retained while the palm target progressively aligns with the selected grasp point.

When a finger-side annotation is available (Appendix C), the palm-alignment objective is

$$\mathcal{L}_{\mathrm{palm}} = \sum_{(t,h)\in\mathcal{J}} \frac{\omega_{t,h}}{2} \sum_{j\in\{\mathrm{f},\mathrm{n}\}} \left[1 - (R^{\mathrm{w}}_{t,h} b^j_h)^\top \bar{a}^j_{t,h}\right], \tag{10}$$

where $\mathcal{J}$ contains valid hand/frame pairs, R^{w} maps wrist-local unit finger and inward-palm axes b^j to the world frame, and $\bar{a}^j$ are their signed unit gripper targets from the selected pose and annotated rail. The fitting weight increases through 20, 100, and 200 over three stages, each using 50 Adam iterations with learning rate 10^{-2}. Positional fitting uses robust joint and contact-center residuals with coefficient 30,000. The effective pose-prior multipliers are 4, 1, and 0.04, and the angle-prior multipliers are 25, 0.01, and 0.01, respectively. Orientation constraints ramp in over five frames, remain fully active from five frames before contact onset through five frames after contact offset, and then ramp out over five frames. Interpolation of the fitted keyframes produces the final reference.

A.5 Imitation Learning

This subsection completes the reward and termination specification of Section 3.3. The notation and illustrations in Figure 2c are provided to aid understanding.

Base tracking reward. We retain InterMimic's physics-based tracking formulation (Xu et al., 2025) and multiply it by the phase-conditioned hand factors in Equation 5. The reference trajectories remain fixed during learning.

Phase factors. Suppressing hand and time indices, the factors in Equation 5 are

$$\begin{aligned} r^{\mathrm{app}} &= r^{\mathrm{open}} r^{\mathrm{aopp}} r^{\mathrm{reach}} r^{\mathrm{plane}}_{\mathrm{app}} r^{\mathrm{side}}_{\mathrm{app}} r^{\mathrm{palm}}_{\mathrm{app}}, \\ r^{\mathrm{grasp}} &= r^{\mathrm{count}} r^{\mathrm{opp}} r^{\mathrm{prox}} r^{\mathrm{plane}}_{\mathrm{grasp}} r^{\mathrm{side}}_{\mathrm{grasp}} r^{\mathrm{palm}}_{\mathrm{grasp}}, \\ r^{\mathrm{rel}} &= r^{\mathrm{open}}_{\mathrm{rel}} r^{\mathrm{withdraw}}. \end{aligned} \tag{11}$$

The orientation terms include $r^{\text{plane}}_{\text{app}} r^{\text{side}}_{\text{app}} r^{\text{palm}}_{\text{app}}$, $r^{\text{plane}}_{\text{grasp}} r^{\text{side}}_{\text{grasp}} r^{\text{palm}}_{\text{grasp}}$, and r^{withdraw} while the others are contact labels or positional rewards.

These factors score pre-opening (r^{open}), approach opposition (r^{aopp}), reaching (r^{reach}), finger contact (r^{count}), grasp opposition and proximity (r^{opp}, r^{prox}), orientation (plane, side, and palm factors), and release, respectively. A phase factor equals one outside its phase; individual factors are also neutral outside their activation windows or when their required data are unavailable. Reference contact intervals define the grasp phase. Approach and release each span at most ten frames before and after contact. Overlapping windows are assigned to the nearer contact boundary, with ties assigned to approach.

Figure 2c organizes geometry guidance into three phases: reaching and pre-orientation before contact, opposition and hand alignment during grasping, and opening and withdrawal after release. These factors complement the tracking objective and combine multiplicatively as in Equation 5 and Equation 11.

Gripper geometry. The gripper provides both positional targets and directional references. Its closing axis d runs across the opening between the two fingers. The unit vector a points from the opening toward the back block and is orthogonalized against d. Thus, the incoming direction is $a_{\text{in}} = -a$, whereas the withdrawal direction is $a_{\text{out}} = a$. The *gripper back center* in Figure 2c is the midpoint of the back block; it guides early reaching and should not be confused with the grasp point used for finger proximity.

Two different planes support the guidance. The *opposition mid-plane*, with unit normal n, determines whether the thumb and other fingers lie on opposite sides of the grasp. The plane labeled *vertical plane of the gripper pose* in the figure is instead the gripper's longitudinal plane, $\Pi_{\text{long}} = c^{\text{grip}} + \text{span}\{d, a\}$, with normal $z = (d \times a)/\|d \times a\|$. It contains the closing and approach directions and rotates with the gripper; "vertical" does not refer to world gravity. All these geometric references are stored in object-local coordinates.

Target assignment with Flexibility. We distinguish targets by their roles. The saved anchor c^{first} identifies the reference grasp point for a hand-contact segment. The proximity target c^{prox} is the common point toward which the thumb and fingers are rewarded for moving. The selected gripper center c^{grip} supplies the *insertion target* shown in Figure 2c. Termination and evaluation use targets c^{term} and c^{eval}, respectively. Object-local points are mapped to world coordinates through Equation 9. Evaluation always uses $c^{\text{eval}} = c^{\text{first}}$, independently of runtime target assignment.

During approach and a newly established unilateral contact, $c^{\text{prox}} = c^{\text{first}}$. To enable flexibility for two hands, at the first bilateral overlap, when neither hand already has an assigned role, target assignment depends on the available finger-side annotations. If only one hand is annotated, it retains its saved anchor and the other receives a flexible target. If both are annotated, the hand closer to its own saved anchor retains that anchor, while the other receives a flexible target. If neither is annotated, both receive flexible targets. A flexible target minimizes the balanced distance in Equation 2 over valid candidates for that segment; if no candidate is valid, the saved anchor is retained. This permits alternative two-hand contact configurations. For a hand with a finger-side annotation, candidate poses must agree with the annotated pose within 45° in both the unsigned closing axis and the signed approach axis.

Once assigned, the target remains fixed in object-local coordinates until that hand's contact interval ends. Its partner's release or re-entry does not change the ongoing assignment; only the incoming hand receives a new role. The selected gripper pose is refreshed when contact-stage target assignment changes. Termination targets and eligibility are specified separately below and need not coincide with proximity-reward targets.

Approach guidance: reaching and pre-orientation. The approaching rewards in Figure 2c prepare the hand to reach the grasp region with a suitable opening and orientation. The pre-opening factor r^{open} measures thumb–finger separation along the closing axis, using level-3 link origins.

Omitting the hand index, the pre-opening reward is

$$
\begin{aligned}
w_t^{\text{hand}} &= \min_{i \in \mathcal{F}_4} \left| \left(p_{\text{th},3}^t - p_{i,3}^t\right)^\top d_t^{\text{world}} \right|, \\
r_t^{\text{open}} &= \exp\left(-\lambda_{\text{open}} \alpha_t^{\text{open}} \left[w_t^{\text{grip}} - w_t^{\text{hand}} \right]_+ \right),
\end{aligned}
\tag{12}
$$

where $\mathcal{F}_4$ contains all four non-thumb fingers, $p_{i,3}^t$ denotes a level-3 link origin rather than a physical fingertip, and $d_t^{\text{world}} = R_t^{\mathcal{O}} d$ is the unit closing axis in world coordinates. The quantity w_t^{grip} is the selected gripper opening expressed in world units. We use $\lambda_{\text{open}} = 5$, while α_t^{open} includes the approach-dependent activation and temporal weighting. The reward penalizes insufficient opening but does not penalize opening wider than the gripper.

The approach-opposition factor r^{aopp} encourages the thumb and the four other fingertips to lie on opposite sides of the opposition mid-plane. Let s_i^{tip} denote fingertip i's signed distance to this plane. We use

$$
Q^{\text{app}} = \frac{1}{4} \sum_{i \in \mathcal{F}_4} \mathbf{1}\left[s_{\text{th}}^{\text{tip}} s_i^{\text{tip}} < 0 \right], \qquad r^{\text{aopp}} = \exp[-\lambda_d \alpha_t^{\text{aopp}} (1 - Q^{\text{app}})],
$$

where α_t^{aopp} controls activation and temporal weighting. Unlike the grasp-opposition reward in Equation 3, this factor uses physical fingertips and does not include an insertion-depth multiplier.

The reaching factor r^{reach} initially brings the hand-proxy center toward the gripper back center. The hand-proxy center is the mean of the wrist and the five first-finger link origins. For forward approaches, frames five and four before contact instead penalize palm-to-grasp-center distance beyond $10\,\text{cm}$. During the final three frames, reaching switches to the balanced thumb–finger distance in Equation 2. Its exponent weight, together with that of approach opposition, increases through 0.5, 0.7, and 0.9. This finger-proximity factor replaces the earlier center-distance factor. The saved grasp point remains fixed during approach, although its associated gripper pose would be reselected by comparing grasp point to wrist direction with the gripper a_{in} direction.

Hand pre-orientation aligns a wrist-fixed axis pointing toward the four-finger side with the annotated side of the gripper opening when finger side annotation is available. Palm-direction guidance also prepares the hand to face into the grasp. These are the directional alignments illustrated in the lower approach and grasp panels; their definitions and activation schedules are given below.

Grasp guidance: opposition, proximity, and insertion. The grasping rewards distinguish three geometric requirements: the fingers should remain near the grasp point, the thumb and other fingers should occupy opposite sides of the mid-plane, and an opposing pair should extend sufficiently into the grasp.

Proximity uses the balanced distance in Equation 2, evaluated at level-2 and level-3 link origins, as labeled in Figure 2c. These are joint-level positions, not physical fingertips. For opposition, let (q_t, n_t) denote the world-space mid-plane and define $s_{i,\ell} = n_t^\top (p_{i,\ell} - q_t)$. A thumb–finger pair is opposing when its signed distances have opposite signs at either matching joint level: $b_i = \max_{\ell \in \{2,3\}} \mathbf{1}[s_{\text{th},\ell} s_{i,\ell} < 0]$.

With a finger-side annotation, the opposition reward in Equation 3 also checks insertion. The depth indicator H equals one only when an opposing pair has both physical fingertips at least $1\,\text{cm}$ past the selected gripper center $F_t(c^{\text{grip}})$ along the incoming direction $R_t^{\mathcal{O}}(-a)$. The same pair must satisfy both opposition and insertion. Proximity excludes the pinky, whereas opposition includes all four non-thumb fingers. Without this annotation, any opposing pair is sufficient ($Q = \max_i b_i$), and the insertion multiplier is omitted to enable finger opposition flexibly in cases in Figure 1.

A separate contact-count factor encourages distributed finger contact: $r^{\text{count}} = \exp[-\lambda_c(1 - \min\{N_{\text{f}}/3, 1\})]$. Here, N_{f} counts fingers with at least one simulated link-contact entry above 0.1. This signal does not by itself identify which object is being touched.

Hand orientation during approach and grasp. The hand-orientation panels in Figure 2c illustrate two signed alignments: which side of the grasp the four fingers face, and which direction the palm faces. Let $R_t^{\mathcal{O}}$ and $R_{t,h}^{\text{w}}$ be the object and wrist rotations. The current world-space hand

directions are $v_j = R^{\mathrm{w}}_{t,h}\bar{v}^j_h$. The inward-palm axis is $\bar{v}^{\mathrm{palm}}_h = (0, 0, -1)$ for both hands; the four-finger-side axes are $(0, 1, 0)$ for the left hand and $(0, -1, 0)$ for the right hand.

The target finger-side direction follows the signed closing axis selected by the annotation. The target palm direction is obtained by projecting the incoming direction onto the opposition mid-plane:

$$v^\star_{\mathrm{palm}} = R^{\mathcal{O}}_t \frac{(I - nn^\top)(-a)}{\|(I - nn^\top)(-a)\|}, \qquad v^\star_{\mathrm{side}} = R^{\mathcal{O}}_t(\sigma d_{\mathrm{anno}}), \tag{13}$$

where d_{anno} is the unit closing axis of the annotated pose, and $\sigma = -1$ or $+1$ specifies the rail assigned to the four-finger side. Equation 4 compares each current direction with its target. Missing annotations or a degenerate target projection make the corresponding signed factor neutral.

An additional, unsigned factor reduces the component of the palm normal perpendicular to the gripper's longitudinal plane (vertical plane):

$$r^{\mathrm{plane}} = \exp\left[-\lambda_d \alpha^{\mathrm{plane}}_t \left|v^\top_{\mathrm{palm}} R^{\mathcal{O}}_t z\right|\right]. \tag{14}$$

This encourages the palm-normal direction to lie within the plane labeled "vertical plane" in the figure. It does not require the palm surface itself to be parallel to that plane. Because it does not choose a signed finger side, this factor remains active with or without a finger-side annotation.

Release guidance: opening and withdrawal. The release rewards in Figure 2c encourage two behaviors: opening the fingers promptly and withdrawing the hand from the object.

For opening, we measure the separation between the thumb and each of the four other fingers along the gripper closing axis. Each pair is measured using both level-3 link origins and physical fingertips, giving eight opening distances. Let $d_j(t)$ be one such distance and $\Delta d_j(t) = d_j(t) - d_j(t-1)$ its increase since the previous frame. The target opening is $0.15\,\mathrm{m}$. On the first release frame, the requested increase $q_j(t)$ closes two-thirds of the remaining gap to this target; on subsequent frames, it closes the full remaining gap, computed from the previous opening. This gradient design is to prevent pre-open fingertips. We penalize insufficient progress toward this target:

$$E_j(t) = \min\left\{1, \frac{[q_j(t) - \Delta d_j(t)]_+}{q_j(t) + \epsilon}\right\}, \qquad r^{\mathrm{open}}_{\mathrm{rel}} = \exp[-5\beta_t \max_{j \in \mathcal{U}_t} E_j(t)], \tag{15}$$

where $\epsilon = 10^{-8}$ prevents division by zero and $\mathcal{U}_t$ contains opening measurements that have not yet reached the target aperture. The error $E_j(t)$ is zero when the requested opening increase is achieved and grows when opening is too slow. Each measurement stops contributing once its target is reached; if all are complete, the reward equals one. The weight β_t is one for the first six release frames, then decreases through 0.8, 0.6, 0.4, and 0.2.

For withdrawal, we compare the actual wrist-motion direction with the outgoing gripper direction a, as illustrated by the withdrawal arrows in Figure 2c. During the first three eligible release frames, $r^{\mathrm{withdraw}} = \exp[-(1 - \cos(\Delta u^{\mathrm{wrist}}_t, a))]$. To measure motion relative to the object, we first express the wrist position in each frame's object-local coordinates, then subtract consecutive positions to obtain $\Delta u^{\mathrm{wrist}}_t$. The reward is highest when the wrist moves along a and remains active even if finger opening has already finished.

Activation schedules and coefficients. The signed palm and finger-side rewards use a $45°$ tolerance and full activation during grasp. Their default weights over the final five approach frames are $(0.1, 0.2, 0.4, 0.6, 0.8)$.

We use $\lambda_c = \lambda_f = 5$ and $\lambda_d = 2$. The signed palm and finger-side penalties also use weight scale 2. Approach opening, reaching, and release opening use scale 5, while approach opposition uses scale 2, before temporal modulation. Distances are measured in meters. These coefficients scale errors inside the exponential factors; the factors themselves are multiplied, not averaged.

Initialization. The training starts from an InterMimic's pretrained policy which is trained in clean RL replay data for each object category, with one enhanced reward for rewarding the index3 finger joint closest to its assigned object vertex (using the same formula and weight for IG reward in InterMimic). Training starts from the first reference frame for 500 epochs for improving locomotion stableness first, then allocates equal reset requests to that frame and validated pre-contact simulated states.

Termination parameters and Geometry-derived Termination. Tracking-based checks inherited from InterMimic monitor root height, human and object tracking, interaction geometry, and required contact. Human/object error thresholds are $0.5/0.8\,\mathrm{m}$. Required-contact loss terminates after more than ten consecutive frames. A 30-frame window additionally detects more than ten frames of forbidden touch or missing required contact.

Our geometry-derived termination (Section 3.3) detects persistent grasp failures: five consecutive evaluable frames without any opposing pair, five consecutive eligible frames with balanced grasp distance above $7\,\mathrm{cm}$, or three consecutive eligible frames with palm-direction error above 100°. Opposition is satisfied by either evaluable hand for retaining flexibility in two hands; proximity and palm checks are applied per hand. Palm checks require a finger-side annotation and valid geometry, and apply during contact and the final three approach frames.

For a new single-hand contact, the termination target depends on whether a finger-side annotation is available. An annotated hand is checked against its saved grasp point, $c^{\mathrm{term}} = c^{\mathrm{first}}$. For an unannotated hand, we select the valid candidate with the smallest balanced distance at each frame. Changing this candidate does not reset the consecutive-failure counter. This selection affects only termination; the proximity reward continues to use the saved grasp point. When two-hand roles are first assigned, every annotated hand is subject to the proximity check. If neither hand is annotated, only the hand closer to its own saved grasp point is checked. For a checked hand, termination uses the same target as its proximity reward, $c^{\mathrm{term}} = c^{\mathrm{prox}}$. Both the selected object-local target and whether the hand is checked remain unchanged until that hand's contact interval ends. If its partner releases or starts a new contact, the ongoing hand keeps its assignment; only the newly contacting hand receives a new one. These rules govern training termination, not the definition of evaluation success. Evaluation follows the separate target-selection and segment-scoring protocol in Appendix B.1.

B Evaluation Details and Additional Results

This section details the evaluation protocol of Section 4 and the **Q5** study.

B.1 Training and Evaluation Protocol

Implementation. We build on InterMimic's SMPL-X-driven humanoid, simulation setup, and physics-based imitation implementation (Xu et al., 2025); our hand guidance, tracking relaxation, initialization, and termination are specified in Appendix A.5. For each sequence, **`InterGuide`**, InterMimic, and RGrailHand use 1,024 parallel environments and 2,000 training epochs, taking approximately 2–3 hours on an NVIDIA GeForce RTX 4090. Multi-sequence training uses 1,024 environments for trashbins and monitors, with 8,000 and 10,000 epochs, respectively. These budgets equal the sums of the corresponding single-sequence epoch budgets ($4 \times 2{,}000$ and $5 \times 2{,}000$) at the same environment count. Chair and table multi-sequence policies instead use 2,048 environments for 10,000 epochs. Thus, **Q5** (Appendix B.2) compares the stated training configurations, not a uniformly compute-matched protocol across all categories.

Reconstruction errors (Q1). Both variants in Table 1 use the same annotated first-grasp points, object-local coordinates, and visibility/validity mask. First-contact error measures hand-to-target distance at each segment's onset. Contact-interval error averages distance to that fixed object-local target throughout the segment. Averaging within segments before averaging across valid segments prevents longer intervals from dominating. Both reconstruction variants use object-local coordinates and a common visibility and validity mask for filtering the occluded frames in the two metrics. These metrics assess contact localization, not complete pose accuracy.

Note that these annotated first-grasp points serve as corrections to the grasp points inferred by the lowest-median-distance criterion, and are used in the subsequent imitation learning. In practice, however, we find that most inferred grasp points are already accurate enough to be used directly, requiring little or no correction.

Success aggregation. Grasp success evaluates finger opposition, grasp proximity, and, when finger-side annotations are available, palm-direction and four-finger-side alignment. The four-finger-

side direction must remain within 60° of its target, and the palm direction within 100°. Opposition and proximity use the frame-wise geometric conditions described in Appendix A.5, rather than the consecutive-failure counters used for training termination. A grasp segment succeeds if there exists a frame within the five-frame window starting from the contact onset, from which all applicable checks remain satisfied through the end of the segment, and the corresponding reference segment is completed. For unannotated two-handed grasp segments, one hand may make non-opposing contact, allowing interactions that combine grasping with support by the other hand. Rollout success requires completion of the reference segment without imposing these additional grasp-quality checks.

Each sequence contributes its fraction of successful segments. Summing these contributions over N sequences gives n, reported as n/N and $100n/N$ percent. Segment outcomes are binary, but sequence contributions can be fractional; each sequence therefore has equal weight regardless of its number of segments. For example, a sequence with two segments of which one succeeds contributes 0.5 to n in the n/N success rate.

Grasp proximity (GProxi). For each hand-contact segment, the saved `first_grasp_index` identifies a fixed object-local target, $c^{\text{eval}} = c^{\text{first}}$. It moves with the simulated object and is shared across methods; neither simulated nearest-point selection nor bilateral reward/termination assignment changes it. Equation 2 evaluates distance to this target over executed reference-labeled grasp-contact frames for both hands, including hands exempt from proximity termination. It uses joint-level link origins rather than physical fingertips and excludes the pinky. This metric measures alignment with the saved grasp point reference.

Contact preservation and tracking. Contact preservation scores only reference-contact frames: a frame is credited when a corresponding hand body reports contact above 0.1. The combined score pools eligible contact frames from both hands. This proxy neither identifies contact with the manipulated object nor establishes grasp success. Body and object errors measure Euclidean root-position drift from the reference in centimeters, not articulation or rotation error. Errors are averaged over executed frames and then equally across sequences.

Rollout selection, baselines and dataset. Table 2 reports the best retained rollout per sequence, selected by length with tracking error as a tie-breaker. The suffix "B" (box rows and InterGuideB) marks multi-sequence training; all other rows use single-sequence training. All methods share InterMimic's policy architecture and fine-tune the same pretrained InterMimic policy weight for each object category.

RGrailHand adds GRAIL-inspired contact-count, opposition, and proximity rewards to InterMimic's imitation learning pipeline. It does not reproduce GRAIL's complete policy structure, learning schedule or data-generation pipeline. It reuses the contact-count factor r^{count} defined in Appendix A.5 and the proximity factor r^{prox} in Equation 2, with $\lambda_c = \lambda_f = 5$. Its opposition factor measures angular separation around the grasp point, rather than placement across the geometry-derived mid-plane. Omitting time and hand indices, let c be the grasp target and use the link origins $p_{i,\ell}$ and finger set $\mathcal{F}_3$ from Equation 2. We define

$$u_{i,\ell} = \frac{p_{i,\ell} - c}{\|p_{i,\ell} - c\|_2 + \epsilon}, \qquad r^{\text{opp,RG}} = \exp\left[-\lambda_d \min_{\ell \in \{2,3\}} \max_{i \in \mathcal{F}_3} \frac{1 + u_{\text{th},\ell}^{\top} u_{i,\ell}}{2}\right], \tag{16}$$

where $\epsilon = 10^{-8}$ and $\lambda_d = 2$. The inner maximum penalizes the least-opposing thumb–finger pair among the index, middle, and ring fingers; the outer minimum selects the better of the two joint levels. The total reward is

$$r_t^{\text{RG}} = r_t^{\text{base}} \prod_{h \in \{L,R\}} r_{t,h}^{\text{count}}\, r_{t,h}^{\text{opp,RG}}\, r_{t,h}^{\text{prox}}, \tag{17}$$

where r_t^{base} denotes the InterMimic's tracking reward. The three hand factors equal one outside reference hand-contact frames; the point-dependent opposition and proximity factors are also neutral when no valid grasp target is available. RGrailHand does not include the additional approach, hand-orientation, or release rewards.

ROrient and GTerm ablations (**Q4**) remove orientation guidance rewards - $r_{\text{app}}^{\text{plane}} r_{\text{app}}^{\text{side}} r_{\text{app}}^{\text{palm}}$, $r_{\text{grasp}}^{\text{plane}} r_{\text{grasp}}^{\text{side}} r_{\text{grasp}}^{\text{palm}}$, and r^{withdraw} for each phase and geometry-derived terminations - opposition, grasp proximity, and palm direction, respectively.

Table 3: **Single- vs. multi-sequence learning (Q5).** InterGuideB trains one policy jointly on all sequences of a category; InterGuide rows repeat Table 2. Training budgets are given in Appendix B.1.

Object type	Method	Success (%,n/N)		Contact preservation (%)	GProxi (cm)	Body track (cm)	Object track (cm)
		Grasp	Rollout				
Chair	InterGuideB	50.0(5.00/10)	50.0(5.00/10)	50.01	25.94	**9.04**	**1.05**
	InterGuide	**87.5(8.75/10)**	**87.5(8.75/10)**	**98.14**	**6.44**	11.27	13.09
Table	InterGuideB	**83.3(5.0/6)**	**83.3(5.0/6)**	**99.70**	**6.18**	**9.46**	**7.98**
	InterGuide	58.3 (3.5/6)	58.3 (3.50/6)	92.06	7.44	11.01	11.38
Monitor	InterGuideB	80.0(4/5)	80.0(4/5)	**99.79**	**5.49**	11.50	14.09
	InterGuide	**100 (5/5)**	**100 (5/5)**	99.27	5.67	**10.60**	**11.93**
Trashbin	InterGuideB	62.5(2.5/4)	75.0(3/4)	98.85	**6.39**	12.18	**10.98**
	InterGuide	**87.5 (3.5/4)**	**100 (4/4)**	**98.95**	6.92	**11.74**	12.02

Evaluation dataset. The six-sequence chair subset is selected for its diversity of contact modes, and every sequence is annotated with the finger side where applicable. Most contain two contact clips, and one contains a single clip. The interactions span pushing, pulling, one- and two-handed rotating, and two-handed pick-up, most of which are shown in the supplementary video. We do not select additional BEHAVE sequences for the imitation evaluation, as they are considerably more difficult: many contain more than five contact clips, and inaccurate scale estimation caused by large initial human–object overlapping and hence collisions.

B.2 Q5: Does Multi-Sequence Training Help?

InterGuideB trains **`InterGuide`** jointly on all sequences of a category with identical rewards, termination, and architecture. Our default multi-sequence configuration achieves 100% rollout success on the box set. For the four grasping categories, Table 3 shows that multi-sequence training improves table grasp and rollout success from 58.3% to 83.3%, whereas single-sequence training achieves higher success on chairs, monitors, and trashbins. Observed table failures frequently involve collisions during intermediate walking, suggesting a potential benefit from sharing locomotion experience across sequences.

The other three sets show a clear downgrade, but for identifiable reasons. On chairs, multi-sequence learning instead suffers from mode collapse across heterogeneous interaction modes with disturbances from monocular reconstruction errors; on monitors and trashbins, it struggles with body balance in some severe cases, whereas single-sequence learning can dedicate more capacity to overcoming it. We view a grasp-point-aware or pose-aware policy and better foot-motion preprocessing as promising ways to make multi-sequence learning more broadly effective on these three sets, while single-sequence learning remains the robust choice when introducing a new interaction mode.

C Interaction Annotation

We collect two lightweight manual inputs with the interface in Figure 5.

Annotate contact labels as contact intervals. To obtain temporally accurate contact labels, we develop a dedicated annotation tool over five contact channels — left hand, right hand, left foot, right foot, and interacting object — illustrated in Figure 5. Given an image sequence, an annotator partitions it into segments and, within each segment, marks the contact intervals of each limb channel either by typing a frame range or by clicking Start and End on the corresponding frames. Whenever a limb contact is annotated, an object contact is created over the same interval automatically, enforcing the physical constraint that limb–object contact implies object involvement and eliminating a common source of annotation inconsistency. The interface supports frame-by-frame browsing, direct jumps to arbitrary frame indices, and on-screen inspection of annotated intervals and current-frame states, which together shorten the labeling loop. Per-segment annotations — cam-

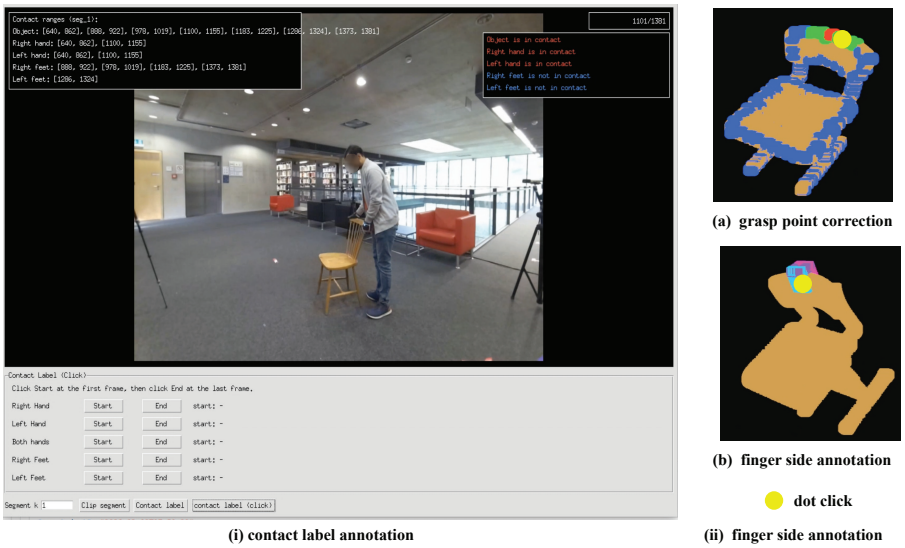

Figure 5: **Annotation interface.** (i) Contact-label annotation is performed before reconstruction: temporal contact intervals are specified independently for the hands and feet, and object contact intervals are inferred from them. (ii) Grasp-point correction and finger-side annotation are performed after reconstruction and before retargeting optionally.

era view, segment boundaries, contact categories, and contact intervals — are serialized to JSON. These intervals define C in Equation 1, the rest phases, and the reward phases.

Optional correction of grasp points and finger side annotation. For grasping, annotators verify or correct the inferred onset grasp center and identify the four-finger rail of the automatically matched DenseGripper pose: the negative or positive closing-axis side, or neither when ambiguous. This resolves the sign ambiguity without manual six-degree-of-freedom pose fitting. A valid center without a rail label still supports positional and opposition guidance, but does not activate signed palm-direction or four-finger-side rewards. Non-grasp interactions require no rail annotation. The verified center becomes c^{first}, and the rail label sets σ in Equation 13.

D Discussion and Future Work

Runtime of reconstruction and retargeting. Reconstruction takes essentially the same time as the original CARI4D pipeline, while grasp-point inference and relative human-scale adaptation add only a few seconds. Retargeting a 10-second sequence takes 150 seconds on average, most of which is spent on keyframe optimization.

Failure cases. We show several failure cases in the demonstration video and outline possible solutions here:

- **F1: Intermediate walking collision.** When manipulating large objects such as tables, the humanoid may collide with the object during the intermediate walking phase; we discuss how multi-sequence learning alleviates this in **Q5**. A pre-processing step that plans collision-free locomotion could also be beneficial.
- **F2: Overly strict termination.** The geometry-derived termination is sometimes too strict, trapping the policy in a local minimum. A curriculum-based exploration schedule could relax this.
- **F3: Residual collisions.** `InterGuide` does not fully prevent humanoid self-collision or collision with the object at release. Release collisions could be addressed on the capture side by enforcing release along a preferred withdrawal direction, or by adding rewards

and termination conditions derived from richer simulation signals to encourage collision avoidance.

Limitations and future work. **`InterGuide`** is currently limited to stable-contact interactions and could be extended to more dynamic cases, such as contact sliding, while still using geometry guidance. Besides, **`InterGuide`** does not yet establish cross-sequence generalization (see also Appendix B.2); additional refinement or distribution-based prior rewards could help. Annotations, particularly contact labels, are still required, though recent vision–language models may enable automating them. There is also room to improve learning efficiency, for example by optimizing simulated object motion to ease transfer and by adopting more efficient sampling to reach longer horizons with multiple contact clips. Finally, transferring these imitated results to physical robots is an appealing next step.