CoWBC: Learning Whole-Body Control for unconstrained Cooperative transport with Humanoid Robots

Anonymous authors
Paper under double-blind review

Abstract

Unconstrained cooperative transport requires humanoid robots to coordinate whole-body actions while maintaining contact with and stable support for a shared object. However, physical coupling through the shared object makes stable whole-body control more challenging than in single-humanoid settings. To address this challenge, we introduce CoWBC, an object-centered framework that learns role-aware cooperative transport policies from real-world human collaboration demonstrations, where roles describe the robots' complementary contributions to the shared transport task. CoWBC uses a shared base policy to learn common cooperative skills and role-specific residuals to adapt actions to each robot's role. Specifically, we first reconstructs interactions between human collaborators and a shared object from synchronized multi-view videos and retargets them into coordinated robot reference motions, preserving the collaborators' relative positions, end-effector–object geometry, and temporal alignment. We then progressively train a centralized teacher with cooperation-aware rewards, then distill its behavior into local student policies. To mitigate the state distribution shift induced by closed-loop student execution, we further augment distillation with lightweight learner-state refinement, where the centralized teacher provides additional supervision on states visited by the current student. This refinement improves closed-loop execution in our evaluations. Extensive experiments show that our dataset achieves a mean valid wrist contact rate 41.10 percentage points higher than that of the evaluated CORE4D subset. CoWBC achieves a safe success rate of 80.18%. Our project website showcases the workflow and visualization results, making them accessible to the community for future research on humanoid cooperation.

1 Introduction

Collaborative transport is a representative task in humanoid cooperation. Humans naturally adapt their motions to both the object and their partners during transport, whereas achieving such behavior with humanoid robots is substantially more challenging in a *unconstrained cooperative* setting, in which each robot must autonomously coordinate their whole-body actions across diverse object motions while maintaining effective contact with and stable support of the same object. Prior work has demonstrated collaborative transport under specific interaction designs (Du et al., 2026; Gao et al., 2024; Lionar & Lee, 2026; Yao et al., 2026), but such unconstrained cooperative transport for humanoid robots remains an open issue in the literature.

Acquiring high-quality demonstrations suitable for unconstrained cooperation is difficult. Capture errors and embodiment differences between humans and robots can prevent retargeted motions from preserving the contact and coordination exhibited in the original demonstrations, which requires a careful balance between processing cost and the reliable preservation of cooperative relationships.

Efficiently transferring unconstrained cooperative priors into policies is challenging. Training must extract task-critical coordination from limited demonstrations while also enabling policies to recover when execution deviates from those priors, which makes it difficult to translate the coordination encoded in demonstrations into policy behavior.

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.

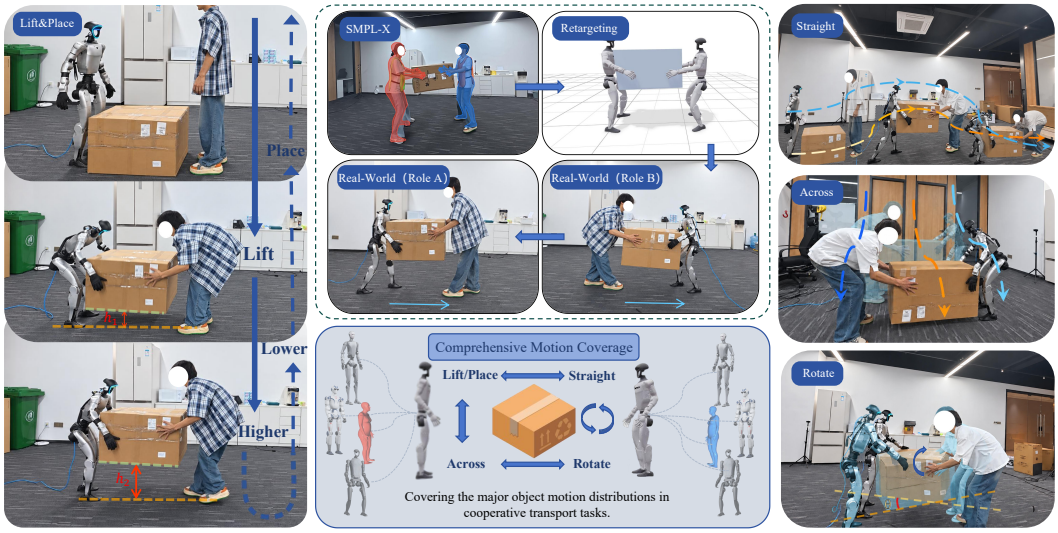


Figure 1: **Unconstrained cooperative transport with flexible motion and agent deployment.** The **left** shows the basic lift-and-place motion, while the **right** presents translation, lateral movement, and rotation. The **middle-top** shows the pipeline from SMPL-X demonstrations through retargeting to real-world deployment. The **middle-center** enables robot deployment in either Role A or B, while the **middle-bottom** supports humans and diverse robot embodiments as cooperative agents.

Meanwhile, unconstrained cooperation places greater demands on policy design. Cooperative transport requires each policy to jointly account for its own state, object motion, and role-specific requirements (i.e., the motion and contact requirements arising from each robot's complementary contribution to the shared transport task) under limited observations while producing mutually compatible whole-body actions, which increases the difficulty of both model design and stable training.

Cooperation learned in simulation is difficult to execute reliably on real robots. Perception errors, object properties, and partner behavior in the real world may differ from the training conditions, which can disrupt the learned coordination and make it difficult for both robots to maintain individual balance and the stability of the shared load.

To address these challenges, we introduce **CoWBC**, a simple yet effective baseline framework for unconstrained cooperative transport with two humanoid robots(Figure 1). At its core, CoWBC adopts an **object-centric shared-base policy with role-specific residuals**. The shared base policy captures common skills required by both robots around the same object, while the role-specific residuals adapt each robot's actions to its own role, which allows the two robots to accommodate changes in object motion while maintaining overall coordination.

In practice, we first train a centralized expert policy that jointly learns individual stability, effective contact, and the shared object objective. Guided by a readiness-gated cooperative reward, training progresses from stable approach through coupled manipulation to robust execution before the expert's behavior is distilled into local policies. Since successful Teacher trajectories do not fully cover the states induced by closed-loop Student execution, we additionally collect Student-visited states and query the centralized Teacher for corresponding supervision. A small fraction of these learner-state samples is incorporated during subsequent distillation, improving robustness to execution-induced distribution shift.

Given that the current data is insufficient to effectively support the training of the models, we further construct a demonstration dataset for dyadic shared-object transport. We reconstruct paired human cooperative motions from multi-view videos and generate paired robot references within a common coordinate system and temporal reference. We define six motion families for cooperative transport. During retargeting, we preserve the relative placement of the two participants and the end-effector positions with respect to the shared object. We then apply relational-consistency and kinematic-

quality checks to select valid synchronized demonstration sequences, which provide coordination priors for subsequent cooperative learning.

Our contributions are threefold:

1. **CoWBC for unconstrained cooperative transport.** We study an unconstrained transport setting in which humanoid robots actively interact with the same shared object. CoWBC provides an end-to-end pipeline spanning paired demonstration construction, cooperative training, and role-decomposed policy learning. Experiments demonstrate effective cooperative transport in simulation and human–robot collaboration on hardware.
2. **Shared and role-specific cooperative policy learning.** We combine a shared base policy with role-specific residuals to capture common cooperative skills while accommodating each robot's complementary role, and train decentralized policies with centralized value information. Cooperation-aware rewards and progressive Teacher training establish physically executable cooperation, which is transferred to local Student policies through distillation and lightweight learner-state refinement.
3. **A paired-demonstration pipeline with diverse motion coverage.** We construct paired demonstrations that comprehensively cover the major object-motion patterns in cooperative transport. Our pipeline preserves contact relationships, relative spatial configurations, and temporal alignment throughout capture, reconstruction, and retargeting, providing robot-compatible priors that retain the demonstrated coordination.

2 Related Work

2.1 Demonstrations and retargeting for cooperative control

Physics-based cooperative human–object interaction (HOI) control draws on individual and paired demonstrations (Xu et al., 2024). CooHOI transfers single-agent transport skills to cooperation through shared-object dynamics feedback (Gao et al., 2024). TeamHOI combines single-human and masked adversarial motion priors with teammate tokens for varying team sizes (Lionar & Lee, 2026). SynAgent integrates solo-to-cooperative transfer, paired interactions, and multi-teacher distillation for trajectory-conditioned control (Yao et al., 2026).

Paired datasets capture shared-object coordination (Kogashi et al., 2026). CORE4D records human–object–human interactions, retargets them to new objects using contact guidance, and demonstrates simulated single-H1 box lifting using individual clips split from paired recordings (Liu et al., 2025a). Harmanoid addresses direct humanoid–humanoid contact through contact-aware retargeting and interaction-specific rewards (Liu et al., 2025b).

Splitting paired demonstrations into individual motions can lose cross-agent coordination. CoWBC preserves relative geometry, shared-object contacts, and temporal alignment during retargeting, improving contact coverage and reference usability for dual-humanoid transport.

2.2 Teacher–student learning and corrective supervision

Teacher–student learning converts motion references into control skills (He et al., 2025). InterMimic combines subject-specific teachers, distillation, and reinforcement learning (Xu et al., 2025), while SynAgent uses multiple teachers and progressive DAgger-style training for cooperative control (Yao et al., 2026). Student-Informed Teacher Training addresses observation asymmetry through joint teacher–student training that encourages imitable behavior (Messikommer et al., 2025).

Beyond observation asymmetry, closed-loop execution can induce distribution shift. DAgger queries experts at learner-visited states (Ross et al., 2011), while LOLS evaluates alternative actions through rollouts (Chang et al., 2015). Q-filter selects demonstration actions by value (Nair et al., 2018), and TS2C gates teacher intervention using value and uncertainty (Xue et al., 2023). Decentralized control requires correcting distribution shift without disrupting coordination. CoWBC addresses this by augmenting Teacher-success distillation with limited supervision at Student-visited states.

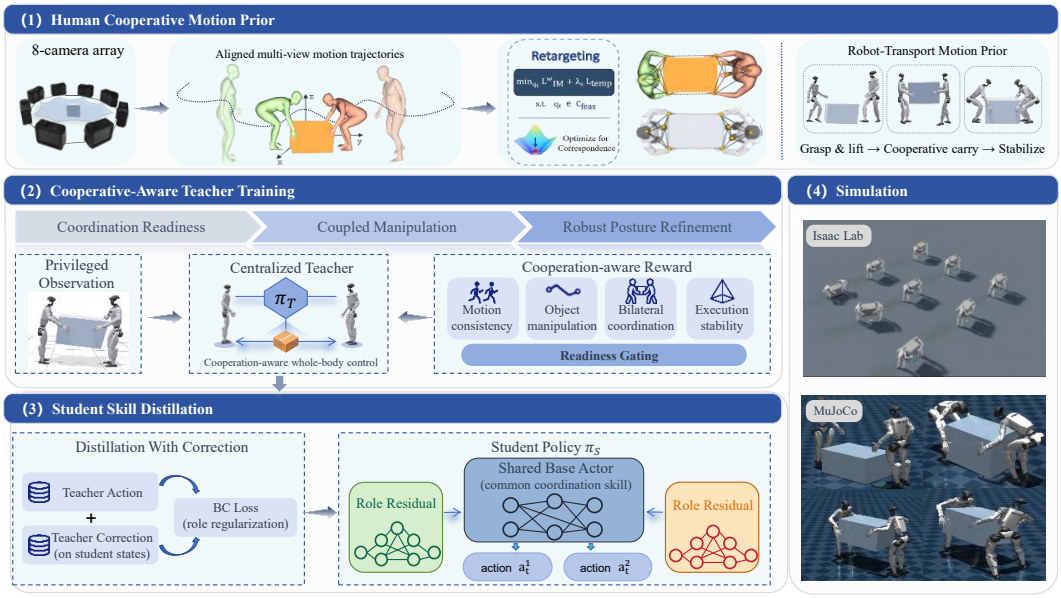


Figure 2: Overview of CoWBC for cooperative humanoid transport. (1) Synchronized multi-view reconstruction and retargeting convert paired human–object demonstrations into robot-transport motion priors. (2) A centralized teacher learns cooperative whole-body control from privileged observations with cooperation-aware rewards and readiness-gated progressive training. (3) Teacher actions and beneficial corrections are distilled into decentralized student policies, combining a shared base actor with role-specific residuals. (4) The resulting role policies are evaluated under dual-robot closed-loop execution across IsaacLab and MuJoCo.

2.3 Parameter sharing and role-specific cooperative control

MAPPO trains decentralized actors with centralized value information (Yu et al., 2022), and is combined with parameter sharing in CooHOI and SynAgent (Gao et al., 2024; Yao et al., 2026). In practice, sharing must also accommodate agent-specific behavior. SePS learns sharing groups (Christianos et al., 2021), ROMA learns latent roles (Wang et al., 2020), and CDS decomposes local action values into shared and individual components to balance sharing and diversity (Li et al., 2021). Whereas CDS decomposes action values, CoWBC combines a shared base actor with role-specific residuals to directly adapt continuous whole-body actions, using MAPPO for cooperative optimization.

Execution interfaces further determine how policies observe partner influence. TeamHOI uses explicit teammate information, CooHOI uses shared-object feedback, and COLA uses proprioception-only execution for human–humanoid transport (Lionar & Lee, 2026; Gao et al., 2024; Du et al., 2026). CoWBC conditions each actor on local proprioception, object observations and history, and object goals and plans, supporting role-specific coordination without direct access to partner states or actions.

3 Methodology

We study cooperative transport with two robots using synchronized human–human–object demonstrations, as illustrated in Figure 2. The participating robots jointly establish contact with, lift, transport, and place a rigid object while maintaining individual balance and shared-load stability. Each robot is assigned one of two roles before execution, which remains fixed within an episode although either robot can assume either role.

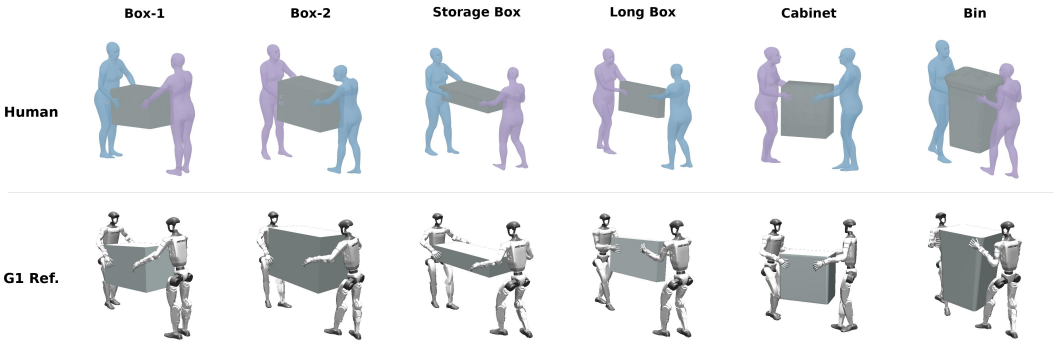


Figure 3: Multi-object cooperative transport demonstrations and their retargeted motion references.

At time t, let x_t^i denote the whole-body state of robot $i \in \{1, 2\}$, x_t^o the manipulated object's state, and $a_t^i \in [-1, 1]^{29}$ robot i's normalized joint action. The coupled system evolves according to

$$s_{t+1} = F(s_t, a_t^1, a_t^2; \xi, d_t), \qquad s_t = (x_t^1, x_t^2, x_t^o), \tag{1}$$

where ξ contains robot, object, and contact-dynamics parameters, and d_t denotes external disturbances at time t. Since both robots act on the same object, either robot's action or tracking error changes the other's interaction conditions.

Given an object-level command $c_t = (g^o, \mathcal{P}_t^o)$ comprising a terminal object goal and short-horizon object-motion plan, robot i computes its action as

$$a_t^i = \pi_S^{r_i}(o_t^i, h_t^{o,i}), \tag{2}$$

where o_t^i contains robot-local proprioception and object/task information, including c_t; $h_t^{o,i}$ contains recent object observations; and r_i indexes the assigned role-specific policy branch.

Section 3.1 constructs paired robot motion priors through coordination-preserving retargeting. Section 3.2 trains and distills a centralized Teacher with learner-state refinement, while Section 3.3 details the object-centric Student's shared base and role-specific residuals.

3.1 Paired Demonstrations for Cooperative Transport

Task coverage. We collect paired demonstrations across six motion families: horizontal translation, yaw rotation, height variation, and their three pairwise combinations. Using large or elongated containers, we vary motion direction, extent, speed, timing, participant pairs, and role assignments. Combined rotation and height variation is included only when safe and repeatable. These motions extend coverage beyond straight transport while preserving relative placement, contact geometry, and temporal coordination.

Synchronized reconstruction. Eight spatially and temporally calibrated cameras record both participants and the shared object to reconstruct SMPL-X bodies and the object's 6D pose. All sequences retain a common world frame and timestamps through approach, contact establishment, lifting, transport, and release. Both participants' pelvis poses and wrist positions are expressed in the object frame to capture relative placement and bilateral manipulation geometry alongside object motion.

Coordination-preserving retargeting. We retarget paired sequences to two G1 robots, mapping roots and end effectors in the shared object frame to preserve bilateral geometry and framewise synchronization (Figure 3). Each reference contains both robots' joint and rigid-body states and the object's pose and velocities.

3.2 COOPERATION-AWARE TEACHER TRAINING AND COOPERATIVE SKILL DISTILLATION

Retargeted paired motions preserve cooperative geometry and timing, but they are not necessarily dynamically executable. We therefore train a centralized Teacher to realize the paired coordination prior under contact, balance, and shared-object constraints, and then distill its behavior into decentralized Students.

Cooperation-aware Teacher learning. The Teacher receives the states, previous actions, and motion references of both robots, together with the object state and task plan, and outputs normalized joint actions for both robots. The Teacher reward groups the training objectives into motion consistency, object manipulation, cooperative relations, and execution stability. The manipulation-dependent terms are softly gated by body-tracking and bilateral-manipulation readiness. Consequently, object progress receives full credit only after both robots establish a dynamically valid cooperative configuration, discouraging unilateral or premature object motion. Detailed reward terms and gates are given in Appendix C.

Progressive objective expansion. Because cooperative transport contains ordered physical prerequisites (stable positioning before lifting, followed by carrying and placement), we train the same Teacher through three successive stages. First, to prepare both robots to initiate transport together, we train on the pre-lifting reference segment, emphasizing motion tracking, balance, and relative positioning so that both robots can reach the transport starting configuration and remain stable there. Second, to learn the complete transport task, we extend training to the full approach–lift–carry–place sequence and activate the object-manipulation and coordination rewards described above. Third, to improve posture and robustness, we continue training with additional arm and wrist posture rewards, stronger stability regularization, broader variations in robot and object dynamics, and external pushes.

Student skill distillation. We distill the trained Teacher's cooperative transport skills into decentralized Students using local observations. A shared base captures common control skills, while role-specific residuals accommodate differences in grasp geometry and load distribution.

From successful Teacher rollouts, we build $\mathcal{D}_T$ by pairing each robot's local observation with its Teacher action. Phase-balanced sampling prevents long stationary segments from overwhelming short, contact-critical transitions. We train the Students with

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{(o,a_T)\sim\mathcal{D}_T}\left[\|\pi_S(o) - a_T\|_2^2\right] + \lambda_{\text{res}}\mathcal{L}_{\text{res}} + \lambda_{\text{role}}\mathcal{L}_{\text{role}}. \tag{3}$$

where $\mathcal{L}_{\text{res}}$ limits residual magnitude and $\mathcal{L}_{\text{role}}$ encourages role-dependent action matching. Because these data cover only Teacher-visited states, we next refine the Students on states encountered during their own execution.

Learner-state refinement. Teacher-success distillation primarily covers states visited by successful Teacher execution, whereas small closed-loop Student errors can induce a different state distribution. We therefore roll out the current Student, query the centralized Teacher at the visited states, and store the resulting local-observation–Teacher-action pairs in a learner-state replay $\mathcal{D}_L$. During a short refinement stage, we optimize the same distillation objective using samples drawn from

$$p_{\text{mix}} = (1-\rho)p_T + \rho p_L, \qquad \rho \ll 1, \tag{4}$$

where p_T and p_L denote the sampling distributions over the Teacher-success replay $\mathcal{D}_T$ and learner-state replay $\mathcal{D}_L$, respectively. Keeping ρ small preserves the nominal cooperative skill while providing limited supervision on Student-induced states.

3.3 OBJECT-CENTRIC ROLE-DECOMPOSED STUDENT POLICY

The centralized Teacher relies on privileged dual-robot observations that are unavailable at execution time. We therefore deploy decentralized Student policies using robot-local proprioception and

Table 1: Comparison of physical quality, cooperative quality, and downstream performance. For the data-quality metrics, each entry reports the sequence-level macro mean / median. RSR reports the overall retargeting success rate. ↓ indicates lower is better, while ↑ indicates higher is better. Best results are in bold.

	Physical quality				Cooperative quality		Downstream
Dataset	**CRS ↓**	**P95-PD ↓**	**FS ↓**	**FGV ↓**	**TCOS ↓**	**VWCR ↑**	**RSR ↑**
CORE4D	**8.50 / 5.71**	3.75 / 3.14	62.98 / 63.29	56.29 / 56.86	23.88 / 21.26	34.59 / 31.21	67.77
Ours	8.75 / 8.15	**3.40 / 2.78**	**41.43 / 42.15**	**30.71 / 29.07**	**22.33 / 18.71**	**75.69 / 92.09**	**87.10**

CRS and P95-PD are reported in cm/s and cm, respectively. CRS and P95-PD are reported in cm/s and cm, respectively; FS, FGV, TCOS, VWCR, and RSR are reported in %.

object-centric task information. For robot i, the Student receives a 195-dimensional current observation o_t^i, comprising 93-dimensional proprioception and 102-dimensional object/task information, together with a 72-dimensional object-history vector $h_t^{o,i}$. Its actor observes neither teammate proprioception or actions nor robot-motion references. Because both Students observe the current and desired motion of the shared object, the object provides a common task coordinate while exposing the effects of partner interaction through its motion.

The two roles share the requirements of whole-body balance, load support, and object-plan tracking, but differ systematically in grasp geometry, local contact conditions, and load distribution. We therefore factorize each policy into a shared base and a role-specific residual:

$$z_t^i = f_{\theta_b}\big(\mathcal{N}(o_t^i)\big), \quad r_t^i = \tanh f_{\theta_r^i}\big(u_t^i, h_t^{o,i}, \mathrm{sg}(z_t^i)\big), \quad a_{S,t}^i = \tanh\big(z_t^i + \lambda_r r_t^i\big). \tag{5}$$

The shared base captures control structure common to both robots, whereas the residual branches represent structured role-dependent corrections. This parameterization keeps both roles within a common cooperative skill space while permitting asymmetric whole-body actions. Detailed feature definitions, network architectures, normalization, and action conversion are provided in Appendix D.

4 EXPERIMENTS

4.1 PHYSICAL AND COOPERATIVE QUALITY OF THE DEMONSTRATIONS

We select CORE4D as the comparison dataset since its two-person collaborative object transport data are among the closest publicly available datasets to our cooperative transport task.

Physical quality is measured by Foot-Ground Violation (FGV) and Foot Sliding (FS) for foot–ground contact, and 95th Percentile Penetration Depth (P95-PD) and Contact Relative Slip (CRS) for hand–box contact. Cooperative quality is measured by Team-Centric Object Stability (TCOS), which tracks box position relative to the carriers, and Valid Wrist Contact Rate (VWCR), the fraction of the carrying phase with valid wrist–box contact. Downstream performance uses Retargeting Success Rate (RSR) for dual-G1 motion retargeting.

For a task-matched comparison, we evaluate cooperative transport sequences from our dataset against CORE4D-Real sequences with the object category *box* and the interaction label *Collaborate*. Metric definitions and preprocessing details are provided in Appendix A.

As shown in Table 1, our dataset performs better on five of the six data-quality metrics and achieves a higher RSR. The increase in VWCR indicates that valid wrist contact is maintained for a larger proportion of the carrying phase. Overall, these results show better contact coverage and downstream retargeting performance for our dataset on the compared cooperative transport sequences.

4.2 EVALUATION FOR COOPERATIVE HUMANOID TRANSPORT

We evaluate cooperative transport policies and ablations using Dual-eval. For directly comparable runs, we use matched motions, reset states, seeds, and evaluation regimes, and assess each rollout

Table 2: Comparison of adapted cooperative humanoid–object interaction methods. Paired demonstrations are used during training, whereas Peer Obs. and Body Ref. indicate direct execution-time inputs to the policy actor, rather than information used by the low-level action decoder.Given that the source code is not publicly available, the SynAgent-style baseline is a second-stage adaptation based on the paper's implementation, rather than a reproduction of the full pipeline.

	Required information			Evaluation metrics					
Method	**Paired Demo**	**Peer Obs.**	**Body Ref.**	**Safe Succ. ↑**	**Coop. Pass ↑**	**Traj. Comp. ↑**	**Fall Rate ↓**	**Drop Rate ↓**	**Grasp Sync Error (m) ↓**
CooHOI-style adaptation	×	×	×	69.32%	21.01%	55.09%	3.14%	27.52%	0.2409
SynAgent-style adaptation	✓	×	✓	0.91%	4.62%	41.75%	99.09%	21.52%	0.1214
Ours: Local Student	✓	×	×	80.18%	82.60%	86.74%	18.54%	1.66%	0.0487
Centralized Teacher[†]	✓	✓	✓	80.30%	83.80%	84.77%	18%	1.80%	0.0536

Notes. ↑ / ↓: higher/lower is better. ✓/×: the corresponding information is/is not directly included in the actor observation. [†]The Teacher actor receives peer observations and body references.

against the final object pose of its reference motion. Our primary metric is Safe Success Rate, the fraction of episodes that complete a stable bilateral grasp, lift, required transport, and maintained placement without a robot fall, an object drop outside the goal region, or failure termination.

We report Cooperation Pass Rate to assess whether transport is accompanied by sustained bilateral grasping, low grasp synchronization error, and consistent wrist heights. Trajectory Completion measures progress through the reference horizon. Fall Rate and Drop Rate identify safety failures, while Grasp Synchronization Error measures the difference between the two robots' wrist-to-object-anchor distances. The grasp-based measures are geometric proxies for coordination rather than measurements of contact force or load sharing. Full thresholds and statistical procedures are provided in Appendix B.

4.3 Comparison of Task-Adapted Cooperative Transport Methods

We adapt representative approaches from CooHOI (Gao et al., 2024) and SynAgent (Yao et al., 2026) to our two-humanoid box-carrying task. All adaptations use our collected human dyadic demonstrations as their raw data source and are trained under budgets that are as comparable as practicable. CooHOI retains its AMP15 embodiment and learns a motion prior from converted single-person trajectories, and SynAgent-G1 is a paper-based adaptation of its Stage-II human–object–human imitation component. Table 2 identifies each adaptation's embodiment, training input, and execution-time information to clarify the scope of comparison.

As shown in Table 2, the CoWBC local Student achieves a Safe Success Rate of 80.18% and a Cooperation Pass Rate of 82.60%, the highest on both metrics among adaptations evaluated on the full transport task. CoWBC also yields lower Grasp Synchronization Error than the other evaluated methods, indicating that its advantage extends beyond reaching the object goal to maintaining stable bilateral grasping during transport.

4.4 Ablation Studies

We conduct controlled ablations to examine three key design choices in CoWBC: the decomposition of shared and role-specific policy components, Teacher reward design and training schedule, and the closed-loop state coverage during Student distillation.

For policy architecture, we remove role-specific residuals, replace them with a shared residual, or train independent actors; we also remove short-term object history. For the Teacher, we compare decoupled manipulation, ungated cooperation, and readiness-gated cooperation, followed by end-

Table 3: Ablations of the Teacher and Student components in CoWBC. Teacher reward and training rows report direct Teacher evaluations; Student architecture and distillation rows report Student evaluations. All variants use the same Dual-eval protocol, motion set, and evaluation reset seeds.

Method variant	Safe succ. ↑	Coop. pass ↑	Traj. comp. ↑	Fall rate ↓	Drop rate ↓	Grasp sync. error ↓
Full CoWBC	80.18%	82.60%	86.74%	18.54%	1.66%	0.0487 m
Policy architecture						
Shared Base Only	70.86%	63.60%	64.56%	28.10%	3.80%	0.0589 m
Parameter-Matched Shared Residual	68.64%	70.63%	84.77%	28.34%	8.67%	0.0563 m
Independent Actors	65.63%	81.69%	67.77%	33.10%	3.59%	0.0636 m
w/o History	64.56%	73.60%	83.66%	35.08%	4.33%	0.0624 m
Teacher and distillation						
Decoupled manipulation	4.82%	0.00%	57.67%	90.35%	31.14%	0.1509 m
Ungated cooperation	60.53%	79.39%	69.83%	38.60%	8.77%	0.0629 m
Gated cooperation (end-to-end)	71.05%	80.70%	71.57%	28.51%	3.51%	0.0506 m
Gated cooperation (staged)	80.30%	83.80%	84.77%	18.00%	1.80%	0.0536 m
w/o Learner-State Refinement	36.40%	57.89%	66.57%	62.28%	25.88%	0.0943 m
w/o Phase-Balanced Replay	62.72%	71.93%	66.18%	36.40%	7.46%	0.0636 m
Multi-agent optimization						
w/o MAPPO Refinement	76.58%	79.18%	79.76%	23.40%	6.09%	0.0619 m
PS-IPPO instead of MAPPO	75.16%	76.33%	77.46%	21.40%	5.54%	0.0589 m

Notes. Safe success is the primary metric. End-to-end and staged gated cooperation use the same objective but different training schedules. "w/o" means without.

to-end versus staged training under the same gated objective and budget. For Student distillation, we remove learner-state refinement or phase-balanced replay.

As shown in Table 3, the ablations show that role-aware policy design, cooperation-aware staged Teacher training, learner-state refinement, phase-balanced replay, and Online cooperative refinement contribute to reliable decentralized cooperative transport.

4.5 Discussion

Our results support modeling cooperative transport as a coupled whole-body control problem. Preserving cross-agent geometry and temporal alignment improves contact coverage and retargeting, increasing the mean Valid Wrist Contact Rate by 41.10 percentage points and the Retargeting Success Rate from 67.77% to 87.10% over the evaluated CORE4D subset. Teacher ablations further show that cooperative coupling, readiness gating, and staged training are important for converting paired motion priors into executable local policies, with the shared object serving as the coordination interface.

CoWBC currently focuses on cooperative whole-body control under preassigned roles and externally supplied object-level plans. Extending the framework to selective learner-state correction and closed-loop dual-humanoid deployment represents an important direction for future work.

5 Conclusion

We presented CoWBC, an object-centered framework that learns cooperative transport from paired human demonstrations through coordination-preserving retargeting, progressive Teacher training, and distillation into decentralized, role-aware policies. The dataset improves contact coverage and retargeting usability, and the full framework achieves an 80.18% overall Safe Success Rate. Within the evaluated task distribution, these results support object-centered, role-aware humanoid control; selective correction and dual-robot hardware validation remain future work.

AI USE STATEMENT

In the course of conducting this work, the generative AI tools were employed primarily to retrieve relevant literature and prior works. Furthermore, we leveraged such tools for manuscript polishing and grammatical verification, with all outputs subjected to careful scrutiny. Additionally, these tools were utilized for data visualization under our supervision. No AI tools were used to generate ideas, design experiments, analyze data, implement algorithms, or produce results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors who submitted papers have read the Code of Ethics and adhere to its guidelines. This work uses publicly available datasets and newly collected human cooperative transport demonstrations. Particularly, all participants provided informed consent prior to data collection. The learned policies are evaluated within a controlled transport setting; their performance does not establish safety for unsupervised deployment around people.

REPRODUCIBILITY STATEMENT

Full details of the model and methodology are presented across the main manuscript and appendix. Hyperparameters, training setups, and evaluation pipelines are clearly enumerated in the supplementary appendix.

REFERENCES

Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2058–2066. PMLR, 2015.

Filippos Christianos, Georgios Papoudakis, Muhammad A. Rahman, and Stefano V. Albrecht. Scaling multi-agent reinforcement learning with selective parameter sharing. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1989–1998. PMLR, 2021.

Yushi Du, Yixuan Li, Baoxiong Jia, Yutang Lin, Pei Zhou, Wei Liang, Yanchao Yang, and Siyuan Huang. Learning human-humanoid coordination for collaborative object carrying. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026.

Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. CooHOI: Learning cooperative human-object interaction with manipulated object dynamics. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79741–79763, 2024.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1516–1540. PMLR, 2025.

Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. MMHOI: Modeling complex 3d multi-human multi-object interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1512–1521, 2026.

Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3991–4002, 2021.

Stefan Lionar and Gim Hee Lee. TeamHOI: Learning a unified policy for cooperative human-object interactions with any team size. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37121–37132, 2026.

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. CORE4D: A 4d human-object-human interaction dataset for collaborative object REarrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1769–1782, 2025a.

Zuhong Liu, Junhao Ge, Minhao Xiong, Jiahao Gu, Bowei Tang, Wei Jing, and Siheng Chen. It takes two: Learning interactive whole-body control between humanoid robots. *arXiv preprint arXiv:2510.10206*, 2025b.

Nico Messikommer, Jiaxu Xing, Elie Aljalbout, and Davide Scaramuzza. Student-informed teacher training. In *International Conference on Learning Representations*, 2025.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6292–6299. IEEE, 2018. doi: 10.1109/ICRA.2018.8463162.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 627–635. PMLR, 2011.

Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. ROMA: Multi-agent reinforcement learning with emergent roles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9876–9886. PMLR, 2020.

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, Yunhui Liu, Wenjun Zeng, and Xiaokang Yang. Inter-X: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22260–22271, 2024.

Siri 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025.

Zhenghai Xue, Zhenghao Peng, Quanyi Li, Zhihan Liu, and Bolei Zhou. Guarded policy optimization with imperfect online demonstrations. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=O5rKg7IRQIO`.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. OmniRetarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. In *2026 IEEE International Conference on Robotics and Automation*, 2026.

Wei Yao, Haohan Ma, Hongwen Zhang, Yunlian Sun, Liangjun Xing, Zhile Yang, Yuanjun Guo, Yebin Liu, and Jinhui Tang. SynAgent: Generalizable cooperative humanoid manipulation via solo-to-cooperative agent synergy. *arXiv preprint arXiv:2604.18557*, 2026.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624, 2022.

APPENDIX

A RETARGETING AND QUALITY EVALUATION

This section provides implementation details for the human-to-robot retargeting procedure and for the physical, cooperative, and downstream quality metrics introduced in Section 4.1.

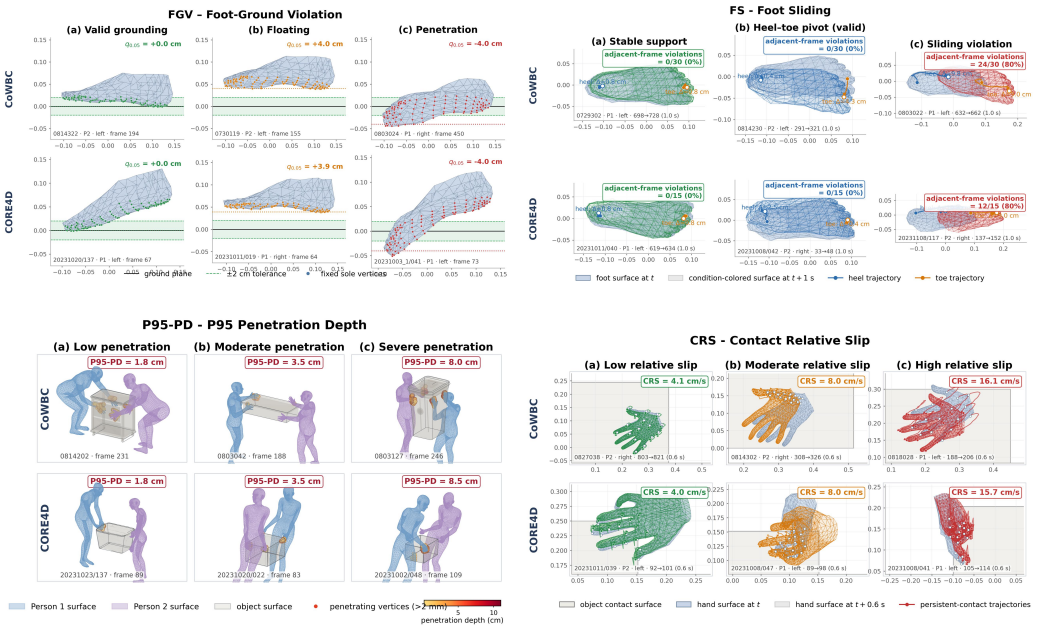


Figure 4: **Physical quality metrics.** Top left: Foot–Ground Violation (FGV), illustrating valid grounding, floating, and ground penetration. Top right: Foot Sliding (FS), illustrating stable support, a valid heel–toe pivot, and a sliding violation. Bottom left: the 95th-percentile Penetration Depth (P95-PD), showing different levels of human–object penetration. Bottom right: Contact Relative Slip (CRS), showing low, moderate, and high hand–object relative slip. The upper and lower rows of each block contain examples from CoWBC and CORE4D, respectively.

For our self-collected demonstrations, each two-person–object interaction is recorded using a spatially and temporally calibrated eight-camera array; the two participants are reconstructed as SMPL-X bodies and the manipulated object as a 6D pose sequence, after which all trajectories are aligned to a common world frame before retargeting.

A.1 Human-to-Robot Retargeting Setup

We adapt the interaction-mesh formulation of OmniRetarget (Yang et al., 2026) to convert synchronized human–human–object demonstrations into paired G1 reference motions. Retargeting uses the shared object trajectory and a common coordinate and temporal frame to retain the spatial organization and timing of the paired interaction. The optimization incorporates joint limits, robot–object and robot–ground non-penetration, support-foot constraints, and temporal smoothing.

A.2 Evaluation Metrics

A.2.1 Physical Quality Metrics

Common evaluation protocol. All four metrics, illustrated in Figure 4 and summarized in Table 4, are first computed for each sequence. Dataset-level macro statistics exclude undefined sequence scores. Specifically, FGV or FS is stored as `NaN` when no valid foot-frame or support transition is available, respectively; CRS is stored as `None` when no valid slip window is available. Distances and speeds are computed internally in meters (m) and meters per second (m/s). P95-PD and CRS are converted to cm and cm/s only for reporting, whereas FGV and FS are reported as percentages. Lower values are better for all four metrics.

FGV: Foot–Ground Violation. FGV measures the fraction of support foot-frames in which the sole either penetrates the ground or floats above it. For each SMPL-X foot, we fix a set of plantar-

surface vertices and define $q_{0.05,h,f,t}$ as their 5th-percentile height along the world vertical axis, where $h \in \{1,2\}$ and $f \in \{L,R\}$. If $|q_{0.05,h,L,t} - q_{0.05,h,R,t}| \le 0.025\,\mathrm{m}$, both feet are labeled as supporting; otherwise, only the lower foot is. We temporally smooth these labels with a majority filter (7 frames for CoWBC at 30 fps and 3 frames for CORE4D at 15 fps). Let $S_{h,f,t}$ be the resulting support mask, $z_g = 0$ the ground height, and $\varepsilon_g = 0.02\,\mathrm{m}$ the valid half-width around the ground. For sequence s,

$$\mathrm{FGV}_s = \frac{\sum_{h\in\{1,2\}} \sum_{f\in\{L,R\}} \sum_{t} S_{h,f,t}\, \mathbb{I}[q_{0.05,h,f,t} < z_g - \varepsilon_g \ \vee \ q_{0.05,h,f,t} > z_g + \varepsilon_g]}{\sum_{h\in\{1,2\}} \sum_{f\in\{L,R\}} \sum_{t} S_{h,f,t}} \times 100\%. \quad (6)$$

FS: Foot Sliding. FS measures excessive horizontal motion of a supporting foot. Unlike FGV, its headline support rule does not require an absolute ground plane: for each participant, the lower of the two sole heights is selected at each frame and the resulting categorical label is temporally smoothed (7 frames at 30 fps and 3 frames at 15 fps). Let $\widetilde{S}_{h,f,t}$ denote this support mask. For $a \in \{\mathrm{heel}, \mathrm{toe}\}$, let $\mathcal{V}_{f,a}$ be the corresponding group of foot-surface vertices. Its horizontal anchor is the component-wise median of the vertices' horizontal coordinates. Only transitions for which $\widetilde{S}_{h,f,t} = \widetilde{S}_{h,f,t+1} = 1$ are evaluated; their index set is denoted by $\mathcal{E}_s$. The horizontal anchor and pivot-aware speed are

$$\begin{aligned} \mathbf{p}^{\mathrm{xy}}_{h,f,a,t} &= \operatorname*{median}_{i\in\mathcal{V}_{f,a}} \mathbf{v}^{h,\mathrm{xy}}_{i,t}, \\ v^{\mathrm{xy}}_{h,f,t} &= \frac{1}{\Delta t} \min_{a\in\{\mathrm{heel,toe}\}} \left\| \mathbf{p}^{\mathrm{xy}}_{h,f,a,t+1} - \mathbf{p}^{\mathrm{xy}}_{h,f,a,t} \right\|_2 . \end{aligned} \quad (7)$$

Taking the smaller heel/toe speed permits a valid pivot about either anchor. With $v_{\mathrm{th}} = 0.075\,\mathrm{m/s}$, the sequence-level violation rate is

$$\mathrm{FS}_s = \frac{\sum_{(h,f,t)\in\mathcal{E}_s} \mathbb{I}\left[v^{\mathrm{xy}}_{h,f,t} > v_{\mathrm{th}} \right]}{|\mathcal{E}_s|} \times 100\%. \quad (8)$$

P95-PD: 95th-Percentile Penetration Depth. P95-PD measures non-physical human–object interpenetration during the automatically detected carrying phase $\mathcal{T}_s$. This phase spans the first to the last sustained object motion (translation above $0.02\,\mathrm{m/s}$ or rotation above $5^\circ/\mathrm{s}$ after temporal smoothing) with $0.5\,\mathrm{s}$ of boundary padding; if no sustained motion is detected, the full sequence is used. Human vertices are transformed into the object-local frame before querying the object SDF; a watertight convex-hull collision proxy is used when an object mesh is not watertight. Using the convention that the SDF is negative inside the object, let $\mathcal{P}_s$ be the multiset of retained positive depths over $h \in \{1,2\}$, $i \in \mathcal{V}_h$, and $t \in \mathcal{T}_s$. The per-vertex depth and sequence score are

$$\begin{aligned} p^h_{i,t} &= \max\left(0, -\mathrm{SDF}_O\left(\mathbf{v}^{h,O}_{i,t}\right)\right), \\ \text{P95-PD}_s &= Q_{0.95}(\mathcal{P}_s). \end{aligned} \quad (9)$$

Here, $\mathcal{V}_h$ is participant h's body-surface vertex set and $Q_{0.95}$ is the empirical 95th percentile. A $0.002\,\mathrm{m}$ numerical tolerance suppresses near-surface noise before pooling but is omitted from the notation for clarity; if no depth remains after this filtering, the score is zero. SDF values and penetration depths remain in meters internally and are converted to centimeters for reporting.

CRS: Contact Relative Slip. CRS measures tangential motion of persistent hand-surface contacts in the object-local frame, removing the object's global translation and rotation. It uses the same sustained object-motion interval as P95-PD but without boundary padding. Contact hysteresis is applied independently to every vertex in the four hand regions: a vertex enters contact when its object-surface distance is at most $0.02\,\mathrm{m}$ and exits only when the distance exceeds $0.03\,\mathrm{m}$. We

Table 4: Definitions and evaluation protocols for physical quality metrics.

Metric	Evaluated Frames	Primitive	Key Threshold	Sequence Aggregation	Unit	↓/↑
FGV	Support foot frames	Sole–ground height	± 2 cm	Violation ratio	%	↓
FS	Consecutive strict-lower support frames	Heel/toe surface anchors	7.5 cm/s	Violation ratio	%	↓
P95-PD	Carrying-phase frames	Body vertices–object SDF	0.002 m tolerance	P95 of pooled depths	cm	↓
CRS	Stable-contact 0.2-s windows	Persistent hand vertices	2/3 cm vertex entry/exit	Mean active-hand slip	cm/s	↓

retain stable episodes with at least five contacted vertices for at least $0.3\,\mathrm{s}$, omit a $0.3\,\mathrm{s}$ boundary margin at each end, and evaluate only $0.2\,\mathrm{s}$ windows containing at least five vertices that remain in contact throughout.

For hand k, vertex i, and valid window beginning at t, let $\mathbf{p}^W_{k,i,t}$ be the world-space position, $(\mathbf{R}^W_{O,t}, \mathbf{t}^W_{O,t})$ the object pose, and $\mathbf{n}^O_{k,i,t}$ the unit object-surface normal at the initial contact. We transform the vertex as $\mathbf{p}^O_{k,i,t} = (\mathbf{R}^W_{O,t})^\top(\mathbf{p}^W_{k,i,t} - \mathbf{t}^W_{O,t})$. With $w = \mathrm{round}(0.2\,\mathrm{s}/\Delta t)$ and $\tau = w\Delta t$, let $\mathcal{V}_{k,t}$ denote the vertices that remain in contact throughout the window, $\Delta\mathbf{p}^O_{k,i,t} = \mathbf{p}^O_{k,i,t+w} - \mathbf{p}^O_{k,i,t}$ their object-local displacement, and $\mathbf{\Pi}_{k,i,t} = \mathbf{I} - \mathbf{n}^O_{k,i,t}\mathbf{n}^{O\,\top}_{k,i,t}$ the tangent-plane projector. We compute

$$\begin{aligned} u_{k,i,t} &= \frac{1}{\tau}\left\|\mathbf{\Pi}_{k,i,t}\Delta\mathbf{p}^O_{k,i,t}\right\|_2, \\ r_{k,t} &= \operatorname*{median}_{i\in\mathcal{V}_{k,t}} u_{k,i,t}. \end{aligned} \tag{10}$$

Let $\mathcal{K}_{s,t}$ be the hands with a valid $r_{k,t}$ and $\mathcal{C}_s$ the set of valid window start frames containing at least one such hand. The sequence score first averages over the active hands in each frame and then over time:

$$\mathrm{CRS}_s = \frac{1}{|\mathcal{C}_s|}\sum_{t\in\mathcal{C}_s}\frac{1}{|\mathcal{K}_{s,t}|}\sum_{k\in\mathcal{K}_{s,t}} r_{k,t}. \tag{11}$$

A.2.2 Cooperative Quality Metrics

TCOS – Team-Centric Object Stability. Team-Centric Object Stability measures variation in the relative configuration of the two participants and the object during collaborative carrying. Let $\mathcal{M}_s$ be the sustained object-motion frames of sequence s. Raw motion is detected when the object's translational speed exceeds $0.02\,\mathrm{m/s}$ or its angular speed exceeds $5^\circ/\mathrm{s}$. The resulting binary mask is majority-smoothed over a $0.3\,\mathrm{s}$ window, and internal gaps no longer than $0.3\,\mathrm{s}$ are filled; sequences with no sustained object motion are excluded. Stable support by a participant requires at least five hand vertices under a $0.02/0.03\,\mathrm{m}$ contact-entry/exit hysteresis for at least $0.3\,\mathrm{s}$. TCOS is applicable only when simultaneous stable support by both participants covers at least 50% of $\mathcal{M}_s$. Otherwise, TCOS and TCOS-Ang are stored as `None` and excluded from dataset-level macro statistics.

For an applicable sequence, TCOS is evaluated on all geometrically valid frames in $\mathcal{M}_s$, rather than only its dual-support frames. We denote this set by $\mathcal{T}_s^{\mathrm{TCOS}}$ and exclude frames with a horizontal pelvis separation below $0.2\,\mathrm{m}$ or a degenerate projected object tangent. If no valid frame remains, the sequence is excluded from dataset-level macro statistics. After applying a $0.2\,\mathrm{s}$ centered moving average, let $\widetilde{\mathbf{p}}_1(t)$, $\widetilde{\mathbf{p}}_2(t)$, and $\widetilde{\mathbf{o}}(t)$ be the two-dimensional horizontal projections of the two pelvis centers and the object center. The team center, participant separation, and top-down axes are

$$\begin{aligned} \mathbf{m}(t) &= \frac{\widetilde{\mathbf{p}}_1(t) + \widetilde{\mathbf{p}}_2(t)}{2}, \qquad d(t) = \|\widetilde{\mathbf{p}}_2(t) - \widetilde{\mathbf{p}}_1(t)\|_2 , \\ \mathbf{x}_h(t) &= \frac{\widetilde{\mathbf{p}}_2(t) - \widetilde{\mathbf{p}}_1(t)}{d(t)}, \quad \mathbf{y}_h(t) = [-x_{h,2}(t) \quad x_{h,1}(t)]^\top . \end{aligned} \tag{12}$$

Thus, $\mathbf{y}_h(t)$ is obtained by rotating $\mathbf{x}_h(t)$ by 90° in the horizontal plane. The object position in this team-centric frame is

$$\mathbf{v}(t) = \begin{bmatrix} (\widetilde{\mathbf{o}}(t) - \mathbf{m}(t))^\top \mathbf{x}_h(t) \\ (\widetilde{\mathbf{o}}(t) - \mathbf{m}(t))^\top \mathbf{y}_h(t) \end{bmatrix} \tag{13}$$

Let D_0 be the object extent along the inferred grasp-face tangent axis. The reference position and separation are the arithmetic means over $\mathcal{T}_s^{\mathrm{TCOS}}$, and the normalized frame-wise errors are

$$\begin{aligned} \bar{\mathbf{v}} &= \frac{1}{|\mathcal{T}_s^{\mathrm{TCOS}}|} \sum_{t \in \mathcal{T}_s^{\mathrm{TCOS}}} \mathbf{v}(t), \qquad \bar{d} = \frac{1}{|\mathcal{T}_s^{\mathrm{TCOS}}|} \sum_{t \in \mathcal{T}_s^{\mathrm{TCOS}}} d(t), \\ e_{\mathrm{pos}}(t) &= \frac{\|\mathbf{v}(t) - \bar{\mathbf{v}}\|_2}{D_0}, \qquad e_{\mathrm{dist}}(t) = \frac{|d(t) - \bar{d}|}{D_0}. \end{aligned} \tag{14}$$

$$\mathrm{TCOS}_s = \frac{1}{|\mathcal{T}_s^{\mathrm{TCOS}}|} \sum_{t \in \mathcal{T}_s^{\mathrm{TCOS}}} \sqrt{e_{\mathrm{pos}}(t)^2 + e_{\mathrm{dist}}(t)^2} \times 100\%. \tag{15}$$

TCOS therefore combines object drift relative to the team center with variation in participant separation. The E_{center} term in Fig. 5 corresponds to e_{pos}.

We additionally report angular stability, $\mathrm{TCOS\text{-}Ang}$. Let $\mathbf{a}^O$ be the inferred object-local grasp-face tangent axis. We rotate it to the world frame, project it horizontally, and normalize it to obtain $\widetilde{\mathbf{a}}(t)$. Its angle in the team-centric frame is $\theta(t) = \mathrm{atan2}(\widetilde{\mathbf{a}}(t)^\top \mathbf{y}_h(t), \widetilde{\mathbf{a}}(t)^\top \mathbf{x}_h(t))$. Because this axis has no forward/backward direction, θ is π-periodic. Its axial mean, wrapped residual, and mean absolute angular deviation are

$$\begin{aligned} \bar{\theta} &= \frac{1}{2} \,\mathrm{atan2}\left(\sum_{t \in \mathcal{T}_s^{\mathrm{TCOS}}} \sin 2\theta(t), \sum_{t \in \mathcal{T}_s^{\mathrm{TCOS}}} \cos 2\theta(t) \right), \\ \delta_\pi(t) &= \frac{1}{2} \,\mathrm{atan2}\big(\sin 2(\theta(t) - \bar{\theta}), \cos 2(\theta(t) - \bar{\theta})\big) , \\ \mathrm{TCOS\text{-}Ang}_s &= \frac{180^\circ}{\pi |\mathcal{T}_s^{\mathrm{TCOS}}|} \sum_{t \in \mathcal{T}_s^{\mathrm{TCOS}}} |\delta_\pi(t)|. \end{aligned} \tag{16}$$

In the experiment, the sequence-level macro-averages and medians of TCOS for CoWBC and CORE4D were $22.33\%/18.71\%$ and $23.88\%/21.26\%$, with corresponding $\mathrm{TCOS\text{-}Ang}$ values of $3.69^\circ/2.76^\circ$ and $5.61^\circ/3.62^\circ$, respectively. CoWBC achieved lower errors for both metrics, indicating that the position and orientation of the object relative to the two-person formation were generally more stable during the collaborative carrying process.

VWCR – Valid Wrist Contact Rate. Valid Wrist Contact Rate measures the fraction of expected participant wrists that maintain plausible load-bearing contact while the object is stably lifted(Figure 6). Let $\mathcal{T}_s^{\mathrm{lift}}$ denote the stable lifted phase of sequence s. We detect this phase from the object-bottom trajectory after a $0.2\,\mathrm{s}$ centered moving average, relative to its fifth-percentile height within the sequence. Lifting uses an entry/exit hysteresis of $0.02/0.01\,\mathrm{m}$, and only episodes lasting at least $0.3\,\mathrm{s}$ are retained. If no such episode exists, the sequence is excluded from dataset-level macro statistics. Candidate load-bearing faces are inferred from the carrier–object geometry near lift onset.

Unlike CRS, which uses the entire hand surface, VWCR uses a predefined wrist/forearm surface region for each wrist. In the neutral SMPL-X template, this region contains proximal-hand and distal-forearm vertices within $0.08\,\mathrm{m}$ of the wrist joint: 221 vertices on the left and 216 on the right.

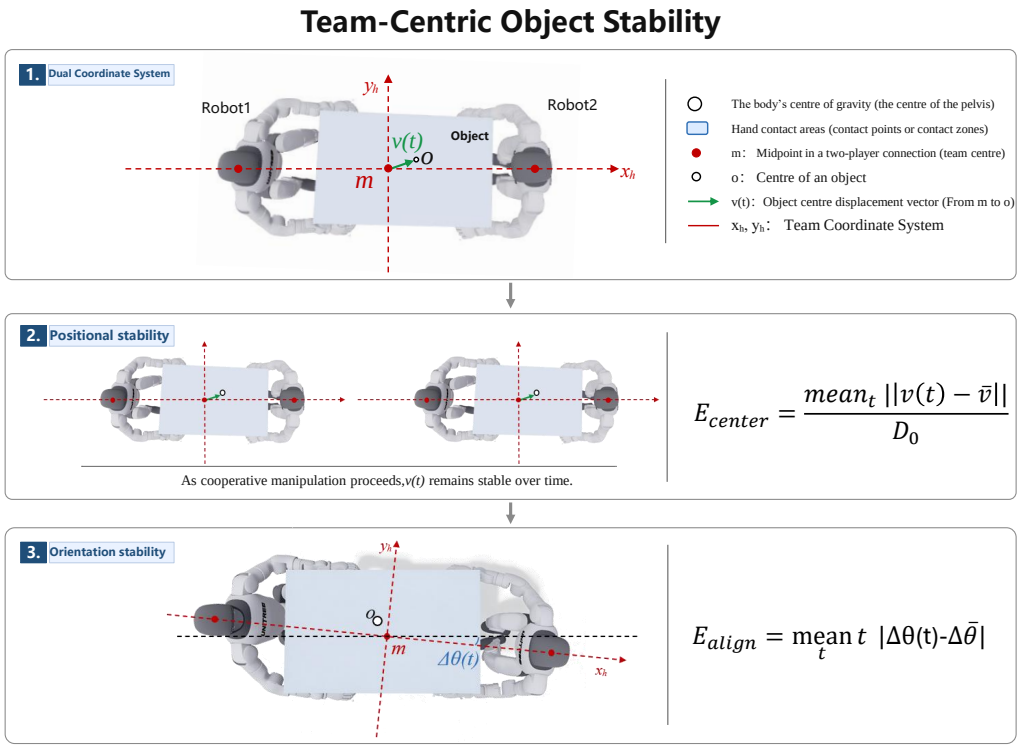


Figure 5: TCOS – Team-Centric Object Stability

For wrist h, let $c_h(t) \in \{0, 1\}$ denote its effective-contact state. We set $c_h(t) = 1$ when at least five region vertices satisfy the load-bearing-face contact test, no selected vertex exceeds $0.02\,\mathrm{m}$ inward penetration, the region's 95th-percentile penetration does not exceed $0.05\,\mathrm{m}$, and the state persists for at least $0.3\,\mathrm{s}$. Vertex contact uses entry/exit thresholds of $0.02/0.03\,\mathrm{m}$.

At each lifted frame, active carriers are inferred from stable full-hand contact using the same $0.02/0.03\,\mathrm{m}$ vertex-level hysteresis, five-vertex minimum, and $0.3\,\mathrm{s}$ duration threshold as above. Let $\mathcal{H}_s(t)$ denote the wrists expected from those active carriers. The frame-level score is

$$s_s(t) = \begin{cases} \dfrac{1}{|\mathcal{H}_s(t)|} \displaystyle\sum_{h \in \mathcal{H}_s(t)} c_h(t), & |\mathcal{H}_s(t)| > 0, \\ 0, & |\mathcal{H}_s(t)| = 0. \end{cases} \tag{17}$$

Thus, one active carrier contributes two expected wrists and two active carriers contribute four; a lifted frame with no identifiable carrier receives a score of zero. The sequence-level VWCR is

$$\mathrm{VWCR}_s = \frac{1}{|\mathcal{T}_s^{\mathrm{lift}}|} \sum_{t \in \mathcal{T}_s^{\mathrm{lift}}} s_s(t) \times 100\%. \tag{18}$$

A higher VWCR indicates that a larger fraction of expected wrists sustain plausible contact without severe penetration. The metric does not require the two wrists of a participant to lie on strictly opposite object faces.

A.2.3 Downstream Usability Evaluation Protocol

RSR – Retargeting Success Rate. Let $\mathcal{E}_D$ denote the set of $N_D = |\mathcal{E}_D|$ eligible two-person source sequences from dataset D before retargeting. For participant p in sequence i, we set $a_{i,p} = 1$ only if retargeting completes and produces a complete, finite G1 trajectory spanning the full source sequence.

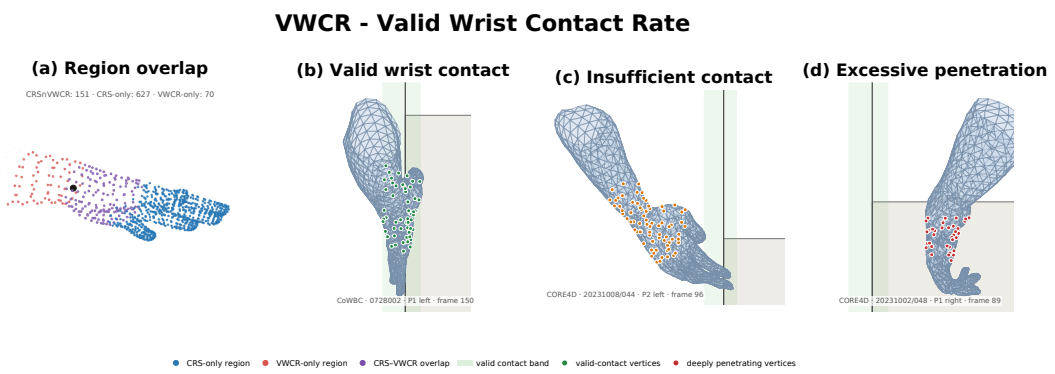


Figure 6: VWCR – Valid Wrist Contact Rate

We then inspect the recombined two-G1–object trajectory using the same predefined manual visual quality-control protocol for both datasets. Let $q_i = 1$ if no gross retargeting artifact that makes the reference unusable is observed, such as discontinuous robot poses, persistent severe robot–object or ground penetration, clear temporal misalignment, or a substantial loss of the intended bilateral hand–object configuration; otherwise, $q_i = 0$. We compute

$$\mathrm{RSR}_D = \frac{1}{N_D} \sum_{i \in \mathcal{E}_D} a_{i,1} a_{i,2} q_i \times 100\%. \tag{19}$$

All failed, incomplete, or visually rejected sequences remain in the denominator. This criterion assesses the usability of the kinematic reference and does not establish dynamic executability.

B Unified Evaluation Framework

We use a single evaluation framework, termed *Dual-eval*, for the centralized Teacher, local Students, comparison methods, and all ablations. The framework separates physical task completion from reference tracking, cooperation quality, and safety. This separation is important because a policy may track a reference without completing the physical task, or may reach the target while relying on asymmetric or unstable interaction. Table 5 summarizes the reported quantities. Safe Success Rate is the only primary success metric; all remaining quantities are diagnostic and are not combined into a second overall score.

B.1 Evaluation Scope and Common Protocol

Each evaluation episode is identified by a motion, evaluation seed, trial index, parallel-environment index, and evaluation regime. These fields define a stable scenario identifier, which is shared across methods for paired comparisons. All compared policies use the same motion set, reset frame, number of trials, number of parallel environments, post-reference horizon, and nominal or robust regime. The protocol file records a version and a hash of the fully resolved configuration. We regard results as directly comparable only when their protocol hashes and scenario identifiers match.

Let $\mathbf{p}_t^o$ and $\mathbf{q}_t^o$ denote the object position and orientation at evaluator step t. The task goal is defined by the final frame of the reference object trajectory,

$$\mathbf{p}^g = \mathbf{p}_{T_{\mathrm{ref}}}^{o,\mathrm{ref}}, \qquad \mathbf{q}^g = \mathbf{q}_{T_{\mathrm{ref}}}^{o,\mathrm{ref}}. \tag{20}$$

The evaluator never substitutes an independently randomized environment goal for this reference-derived goal. Teacher and Student policies therefore face the same physical task definition even though they may receive different policy observations. Pure reference deviation is recorded for tracking diagnostics but is disabled as a default early-termination rule, preventing a tracking-specific termination from changing the physical task horizon of one policy class.

Table 5: Overview of the Dual-eval metrics. Binary episode outcomes are aggregated as rates, whereas continuous quantities are summarized over episodes. The cooperation quantities are geometric proxies and do not measure contact force or physical load sharing.

Axis	Metric	Meaning	Direction
Task completion	**Safe Success Rate**	Stable completion of the target milestone without a fall, object drop, failure termination, or loss of the maintained placement.	↑
Task completion	Grasp, Lift, Transport, Place, and Release Rates	Stage-wise outcomes that localize where an episode succeeds or fails.	↑
Reference tracking	Imitation Pass Rate	Whether reference progress and terminal object/robot tracking errors satisfy the prescribed tolerances.	↑
Reference tracking	Object and Robot Tracking Errors	Position, orientation, velocity, root, lower-body, and joint errors relative to the reference motion.	↓
Cooperation	Cooperation Pass Rate	Whether transport is achieved while bilateral support, grasp synchronization, and wrist-height consistency all pass.	↑
Cooperation	Dual Support Ratio	Fraction of the carry phase for which both robots simultaneously maintain the geometric grasp criterion.	↑
Cooperation	Grasp Synchronization Error	Difference between the two robots' wrist-to-anchor distances; lower values indicate more symmetric grasp evolution.	↓
Cooperation	Wrist Height Difference	Vertical mismatch between the two robots' grasping wrists.	↓
Safety	Fall Rate	Fraction of episodes in which either humanoid falls.	↓
Safety	Drop Rate	Fraction of episodes with an out-of-goal ground drop or an out-of-goal airborne release followed by rapid descent.	↓
Safety	Failure Reason	Mutually exclusive diagnostic category identifying the earliest decisive task or safety failure.	—

The evaluator represents cooperative transport as the ordered state sequence

$$\text{stable bilateral grasp} \rightarrow \text{lift} \rightarrow \text{transport} \rightarrow \text{stable place} \rightarrow \text{optional release}. \tag{21}$$

Later milestones can be reached only after their prerequisites. This ordered construction prevents, for example, an initially elevated object from being counted as a successful lift or an object that starts close to the goal from being credited with transport without prior grasp and lift.

B.2 Task Completion and Safe Success

The task-completion thresholds are summarized in Table 6.

Stable bilateral grasp. For robot $i \in \{1, 2\}$, let d_t^i be the maximum distance from any configured wrist to its reference-derived object-frame grasp anchor. We convert this distance into a bounded per-robot grasp score,

$$s_t^i = \text{clip}\left(\frac{d_{\text{zero}} - d_t^i}{d_{\text{zero}} - d_{\text{full}}}, 0, 1\right), \qquad s_t^{\text{bil}} = \min(s_t^1, s_t^2), \tag{22}$$

where $d_{\text{full}} = 0.04\,\text{m}$ and $d_{\text{zero}} = 0.18\,\text{m}$. A stable bilateral grasp requires $s_t^{\text{bil}} \geq 0.5$ for five consecutive evaluator steps. Taking the maximum wrist distance within each robot and the minimum score across robots ensures that a single nearby wrist cannot satisfy the bilateral criterion. This score is a geometric grasp proxy; it is not a measurement of contact force.

Lift. Let z_0^o be the object's measured height immediately after the episode reset. Lift height is defined relative to this realized initial state,

$$h_t^{\text{lift}} = z_t^o - z_0^o. \tag{23}$$

The lift milestone requires a preceding stable bilateral grasp and $\max_t h_t^{\text{lift}} \geq 0.05\,\text{m}$.

Transport. Let

$$D_{\text{ref}} = \left\| \mathbf{p}_{xy}^g - \mathbf{p}_{0,xy}^o \right\|_2 \tag{24}$$

be the reference-directed horizontal displacement. Motions with $D_{\text{ref}} < 0.08\,\text{m}$ are classified as lift-and-place tasks, for which horizontal transport is not required. Otherwise, with

$$\mathbf{u}^g = \frac{\mathbf{p}_{xy}^g - \mathbf{p}_{0,xy}^o}{D_{\text{ref}}}, \tag{25}$$

the achieved directed progress and required progress are

$$P_t = \max\!\left(0, (\mathbf{p}_{t,xy}^o - \mathbf{p}_{0,xy}^o)^\top \mathbf{u}^g\right), \qquad P_{\text{req}} = \min(0.20\,\text{m},\, 0.5 D_{\text{ref}}). \tag{26}$$

Transport requires a preceding lift and $\max_t P_t \geq P_{\text{req}}$. Directed net displacement is used instead of a decrease in goal distance so that lateral motion or an advantageous initialization cannot be misclassified as transport.

Stable place. Placement is evaluated against the final reference pose and the physical support surface. The object's lowest point is computed from its oriented bounding box rather than from its center height. If $h_{\text{vert}}(\mathbf{q}_t^o)$ is the vertical half-extent induced by the current orientation, the bottom height is

$$z_t^{\text{bottom}} = z_t^o - h_{\text{vert}}(\mathbf{q}_t^o). \tag{27}$$

Let z^{sup} denote the support-surface height, and define

$$e_t^{xy} = \left\| \mathbf{p}_{t,xy}^o - \mathbf{p}_{xy}^g \right\|_2, \tag{28}$$

$$e_t^{\text{sup}} = \left| z_t^{\text{bottom}} - z^{\text{sup}} \right|, \tag{29}$$

$$e_t^R = 2 \arccos(|\langle \mathbf{q}_t^o, \mathbf{q}^g \rangle|). \tag{30}$$

A candidate placement requires

$$e_t^{xy} \leq 0.08\,\text{m}, \quad |e_t^{\text{sup}}| \leq 0.02\,\text{m}, \quad e_t^R \leq 0.35\,\text{rad}, \tag{31}$$

together with object linear speed at most $0.08\,\text{m}\,\text{s}^{-1}$ and angular speed at most $0.35\,\text{rad}\,\text{s}^{-1}$. All conditions must hold for ten consecutive steps to establish a stable placement and for a further twenty steps to establish *maintained placement*. Thus, a suspended object, a transient passage through the goal, or an object that is immediately removed after touchdown is not counted as successful.

Safe success. The experiment uses stable placement as the headline task milestone. For episode e, the primary binary outcome is

$$y_e^{\text{safe}} = y_e^{\text{placed}} \wedge y_e^{\text{maintained}} \wedge \neg y_e^{\text{failure-termination}} \wedge \neg y_e^{\text{fall}} \wedge \neg y_e^{\text{drop}}. \tag{32}$$

Safe Success Rate is the mean of y_e^{safe} across evaluated episodes. The grasp, lift, transport, place, and optional release rates are reported separately to diagnose stage-wise failure, but they do not replace Safe Success Rate as the primary outcome.

B.3 Reference-Trajectory Tracking

Reference tracking is evaluated independently of physical task success. Let $c_e \in [0, 1]$ be the fraction of the reference horizon completed in episode e. The binary imitation outcome requires

$$\begin{aligned} y_e^{\text{imit}} = {} & [c_e \geq 0.95] \wedge [e_{e,T}^{xy} \leq 0.20\,\text{m}] \wedge [e_{e,T}^z \leq 0.12\,\text{m}] \\ & \wedge [e_{e,T}^R \leq 0.60\,\text{rad}] \wedge [e_{e,\max}^{\text{lift-def}} \leq 0.12\,\text{m}] \wedge [e_{e,\max}^{\text{root}} \leq 0.60\,\text{m}] \\ & \wedge \neg y_e^{\text{failure-termination}}, \end{aligned} \tag{33}$$

Table 6: Task-completion thresholds used by the common evaluation protocol. Consecutive-step requirements reject transient contacts and brief passages through the target region.

Criterion	Threshold	Temporal condition
Bilateral grasp score	≥ 0.50	5 consecutive steps
Lift height	$\geq 0.05\,\mathrm{m}$	after stable grasp
Transport progress	$\min(0.20\,\mathrm{m}, 0.5D_{\mathrm{ref}})$	after lift
Final horizontal error	$\leq 0.08\,\mathrm{m}$	10 stable + 20 maintained steps
Support-height error	$\leq 0.02\,\mathrm{m}$	same placement window
Orientation error	$\leq 0.35\,\mathrm{rad}$	same placement window
Object linear speed	$\leq 0.08\,\mathrm{m\,s^{-1}}$	same placement window
Object angular speed	$\leq 0.35\,\mathrm{rad\,s^{-1}}$	same placement window

where $e^{\mathrm{root}}_{e,\max}$ is the maximum root-position error over both robots, and the lift-deficit term is applied when relevant to the policy profile. Imitation Pass Rate is the mean of y^{imit}_e.

For diagnostic analysis, Dual-eval additionally records object position, orientation, and velocity errors; per-robot root errors; lower-body position error; and joint-position root-mean-square error. Their mean, final, maximum, and upper-tail summaries are retained at the episode level. These quantities describe how a behavior differs from the reference, but they do not determine Safe Success. Consequently, a policy may complete the physical transport task without passing the imitation criterion, and a reference-tracking policy may fail Safe Success if it does not achieve a stable physical placement.

B.4 Cooperation Quality

The cooperation axis measures whether the two robots maintain a coordinated geometric relationship through the shared object. Let $\mathcal{C}_e$ be the set of evaluator steps belonging to the carry phase of episode e. The dual support ratio is

$$r^{\mathrm{dual}}_e = \frac{1}{|\mathcal{C}_e|} \sum_{t \in \mathcal{C}_e} \mathbb{I}\big[s^{\mathrm{bil}}_t \geq 0.5\big]\,. \tag{34}$$

For robot i, let $\bar{d}^i_t$ be the mean wrist-to-anchor distance over its configured wrists. The instantaneous grasp synchronization error is

$$e^{\mathrm{sync}}_t = \big|\bar{d}^1_t - \bar{d}^2_t\big|\,. \tag{35}$$

We report its 95th percentile over the carry phase so that a short but substantial asymmetric grasp episode is not hidden by the temporal mean. We similarly compute the 95th percentile of the absolute wrist-height difference, denoted $e^{\mathrm{height},95}_e$.

The binary cooperation outcome is

$$y^{\mathrm{coop}}_e = y^{\mathrm{transported}}_e \wedge [r^{\mathrm{dual}}_e \geq 0.80] \wedge [e^{\mathrm{sync},95}_e \leq 0.08\,\mathrm{m}] \wedge [e^{\mathrm{height},95}_e \leq 0.10\,\mathrm{m}]. \tag{36}$$

Cooperation Pass Rate is used only to interpret behavior and is not conjoined with Safe Success. Because these measurements are based on wrist–anchor geometry, they quantify bilateral grasp maintenance and synchronization, not contact force, partner effort, or payload sharing. Claims about physical load sharing would require force or contact sensing not assumed by the present evaluation protocol.

B.5 Safety and Failure Attribution

Robot fall. An episode is labeled as a fall when the environment terminates and the root height of either robot is below the configured termination height. Fall Rate is reported separately from Safe Success Rate to distinguish failures of whole-body stability from failures of object manipulation.

Object drop. A drop is defined only outside the goal region and only after the object has been lifted. Dual-eval detects two complementary cases:

$$y_t^{\text{drop-ground}} = y_t^{\text{lifted}} \wedge y_t^{\text{outside-goal}} \wedge y_t^{\text{support-contact}}, \tag{37}$$

$$y_t^{\text{drop-air}} = y_t^{\text{lifted}} \wedge y_t^{\text{outside-goal}} \wedge [s_t^1 < s_{\text{release}}] \wedge [s_t^2 < s_{\text{release}}] \wedge [v_{z,t}^o \leq -0.30\,\text{m}\,\text{s}^{-1}], \tag{38}$$

where $s_{\text{release}} = 0.15$. A drop failure is registered when either signal persists for three consecutive steps. Restricting drop detection to the region outside the goal prevents a controlled descent onto the target surface from being counted as failure.

Failure taxonomy. For each unsuccessful episode, the evaluator assigns one primary failure reason using the following fixed priority:

1. robot fall;
2. object drop;
3. no stable bilateral grasp;
4. no lift;
5. no transport, when transport is required;
6. placement position outside tolerance;
7. object not on the support surface;
8. placement orientation outside tolerance;
9. placement speed above tolerance;
10. insufficient stable-placement duration;
11. placement not maintained;
12. no release, when release is required;
13. other termination or incomplete rollout.

The failure label explains why the task was unsuccessful and is distinct from the termination reason, which records how the simulator or evaluator ended the rollout. For example, a rollout may reach the evaluator horizon with `place_not_maintained` as its failure reason.

B.6 Nominal and Robust Evaluation Regimes

We evaluate nominal capability and physical robustness as separate regimes. In the nominal regime, every episode starts from the prescribed reset frame, physical domain randomization is disabled, and no external push is applied. This regime measures task capability under the canonical model parameters.

In the robust regime, reset-time physical randomization is retained and a stable scenario seed determines the realized object and robot parameters. The evaluator records the realized mass, friction, center-of-mass, and motor strength values. Environment-internal random pushes are disabled and replaced with a deterministic evaluator-controlled push schedule, ensuring that paired policies receive the same disturbance for a shared scenario identifier.

Nominal and robust episodes are not paired across regimes. Methods are first compared within each regime using matched scenarios. Robustness degradation is then summarized as

$$\Delta_{\text{rob}} = \text{SSR}_{\text{nominal}} - \text{SSR}_{\text{robust}}, \tag{39}$$

where SSR denotes Safe Success Rate. Reporting both absolute rates and Δ_{rob} prevents a method with weak nominal performance from appearing robust merely because it has little additional performance to lose.

Table 7: Centralized Teacher observation. The per-robot block is expressed in that robot's current base frame where applicable, and the two blocks are concatenated before being passed to the policy.

Component	Dimension per robot	Representation
Base angular velocity	3	local vector
Projected gravity	3	local vector
Relative joint position	29	joint coordinates
Relative joint velocity	29	joint coordinates
Previous action	29	normalized action
Reference root position	3	base frame
Reference root orientation	6	6D rotation, base frame
Reference joint position	29	joint coordinates
Reference joint velocity	29	joint coordinates
Current object position	3	base frame
Current object orientation	6	6D rotation, base frame
Final object position	3	base frame
Ten reference-object waypoints	90	$10 \times (3 + 6)$, base frame
Total	262 per robot; 524 centralized	two robot-centred blocks

C Centralized Teacher-Model Training

This appendix specifies the Teacher-training details omitted from Sec. 3.2. We first define the centralized observation, action, and cooperative objective, and then describe how rollout support, reset distributions, reward activation, and physical randomization are jointly expanded during training. We finally present the controlled ablations used to separate objective design, readiness gating, stage-wise optimization, and robustness continuation. Evaluation criteria are shared by all Teacher, Student, baseline, and ablation experiments and are defined separately in Appendix B.

C.1 Centralized Teacher Interface

Centralized observation and action. The Teacher receives the states and motion references of both robots together with the state and reference plan of the shared object, as summarized in Table 7. Its observation is the concatenation of two 262-dimensional robot-centred blocks. Each block contains 93 dimensions of proprioception, 67 dimensions of robot-motion reference, and 102 dimensions of object and task information, giving a 524-dimensional centralized observation.

For each robot, the proprioceptive block consists of local base angular velocity, projected gravity, joint positions and velocities relative to the default state, and the previous action. The robot-reference block contains the reference root position and 6D orientation in the current base frame, followed by reference joint positions and velocities. The object/task block contains the current object position and 6D orientation, the final reference-object position, and ten reference-object waypoints. Each waypoint contains a three-dimensional position and 6D orientation expressed in the current robot base frame. The waypoint sequence starts from the current reference frame and is sampled at one-frame intervals; indices beyond the end of a motion are clamped to its final frame.

The Teacher produces 29 residual joint commands for each robot, $a_t^T = [a_t^{T,1}, a_t^{T,2}] \in [-1, 1]^{58}$. After clipping, the normalized commands are mapped to reference-centred joint-position targets,

$$q_t^{\mathrm{des},i} = q_t^{\mathrm{ref},i} + m_t^i \odot (q_{\max}^i - q_{\min}^i) \odot a_t^{T,i}, \qquad i \in \{1, 2\}, \tag{40}$$

where m_t^i is the randomized motor-strength multiplier. The targets are executed by joint-space PD controllers. The simulator and policy run at 50 Hz in the current implementation.

Table 8: Stage-wise activation of the main Teacher reward terms. Values are reward coefficients; zero denotes a disabled term. Stability and safety regularizers remain active throughout, with stronger coefficients during final refinement.

Reward term	Approach learning	Coupled manipulation	Robust refinement	Activation
Dual-robot motion tracking	4.0	4.0	4.0	always
Inactive-object motion penalty	-2.0	-2.0	-2.0	$1 - g^{\mathrm{ref}}$
Reference-anchor proximity	0	3.0	2.8	g^{body}
Bilateral anchor synchronization	0	2.5	2.5	$g^{\mathrm{ref}} g^{\mathrm{body}}$
Wrist-height balance	0	1.5	1.5	$g^{\mathrm{ref}} g^{\mathrm{body}}$
Wrist–object velocity match	0	0.75	0.75	$g^{\mathrm{ref}} g^{\mathrm{body}}$
Persistent anchoring	0	1.5	1.5	$g^{\mathrm{ref}} g^{\mathrm{body}}$
Lift progress / object lift	0 / 0	6.0 / 5.0	6.0 / 5.0	readiness gated
Object pose/velocity tracking	0	2.0	2.0	$g^{\mathrm{ref}} g^{\mathrm{body}} g^{\mathrm{coord}}$
Object-drop penalty	0	-8.0	-8.0	$g^{\mathrm{ref}} g^{\mathrm{body}}$
Wrist-reference tracking	0	0	0.15	g^{body}
Arm / wrist posture tracking	0 / 0	0 / 0	0.7 / 1.2	always

C.2 Cooperative Reward Design and Progressive Teacher Training

Cooperative task objective. The task reward is decomposed as

$$r_t = r_t^{\mathrm{track}} + r_t^{\mathrm{object}} + r_t^{\mathrm{coord}} + r_t^{\mathrm{posture}} + r_t^{\mathrm{stability}} + r_t^{\mathrm{reg}}. \tag{41}$$

Here, r_t^{track} measures dual-robot root, key-body, body-velocity, and selected joint tracking; r_t^{object} measures object lift and pose/velocity tracking and penalizes premature motion and dropping; r_t^{coord} measures reference-anchor proximity, bilateral synchronization, wrist-height balance, wrist–object velocity consistency, and persistent anchoring; r_t^{posture} refines arm and wrist posture; and $r_t^{\mathrm{stability}}$ and r_t^{reg} penalize unstable body motion, foot sliding, action variation, joint-limit violations, and failure. The stage-wise activation of these reward terms is summarized in Table 8.

Downstream manipulation rewards are modulated by continuous readiness gates. Let $q_t^i \in [0, 1]$ denote the tracking quality of robot i, computed from its pose, body-orientation, and body-velocity scores. The body-readiness gate is

$$g_t^{\mathrm{body}} = 0.15 + 0.85 \min(q_t^1, q_t^2). \tag{42}$$

Let q_t^{anchor}, q_t^{sync}, and q_t^{height} denote bilateral anchor quality, anchor-distance synchrony, and wrist-height balance. Bilateral carry readiness and its object and lift gates are

$$c_t = 0.55 q_t^{\mathrm{anchor}} + 0.25 q_t^{\mathrm{sync}} + 0.20 q_t^{\mathrm{height}}, \tag{43}$$

$$g_t^{\mathrm{coord}} = 0.25 + 0.75 c_t, \qquad g_t^{\mathrm{lift}} = 0.20 + 0.80 c_t. \tag{44}$$

A separate reference-activity gate g_t^{ref} increases linearly from zero to one when the reference object exhibits sufficient linear motion, angular motion, or vertical displacement. The thresholds/full-activation values are $0.12/0.25\,\mathrm{m\,s^{-1}}$, $0.12/0.35\,\mathrm{rad\,s^{-1}}$, and $0.04/0.08\,\mathrm{m}$, respectively.

These gates are soft and use no hysteresis. Object tracking is weighted by $g_t^{\mathrm{ref}} g_t^{\mathrm{body}} g_t^{\mathrm{coord}}$; lift progress by $g_t^{\mathrm{ref}} g_t^{\mathrm{body}} g_t^{\mathrm{lift}}$; and object lift by $g_t^{\mathrm{ref}} g_t^{\mathrm{body}} g_t^{\mathrm{coord}}$. The drop penalty depends on $g_t^{\mathrm{ref}} g_t^{\mathrm{body}}$ but not on bilateral readiness, so poor coordination cannot suppress a safety-relevant penalty. During reference-static phases, $(1 - g_t^{\mathrm{ref}})$ instead activates the inactive-object penalty. Persistent anchoring is estimated from ten consecutive frames in which all wrists remain within $0.10\,\mathrm{m}$ of their reference-derived object anchors. Because it uses geometric proximity rather than force sensing, we refer to it as *proximity-based persistent anchoring*.

The Teacher is trained by expanding four coupled elements: the reference segment visited by rollouts, the reset distribution within that segment, the active reward subset, and the magnitude of physical randomization. This joint expansion is essential: activating long-horizon object objectives before the policy can repeatedly reach manipulable states produces sparse or misleading gradients, whereas introducing posture and strong disturbances before stable carrying can disrupt task acquisition.

Table 9: Joint expansion of rollout support, reset distribution, reward activation, and physical uncertainty during Teacher training.

Stage	Learning target	Rollout/reset support	Newly active signals	Randomization level
Coordination-ready approach	approach, placement, stopping	approach segment; 0.70 start + 0.30 recovery	motion, stability, inactive-object control	light robot DR
Coupled object manipulation	lift, carry, place, bilateral consistency	full reference; annealed start/early/mid resets	gated coordination, lift, object tracking, drop prevention	robot + object DR
Robust posture refinement	posture, smoothness, disturbance recovery	full reference; same reset curriculum	posture and stronger stability regularization	broad robot/object DR + pushes

Table 10: Domain-randomization ranges used by the three training stages. Robot base mass is additive; motor strength and object mass are multiplicative; COM and joint perturbations are per-axis. Non-degenerate ranges are sampled uniformly.

Parameter	Approach learning	Coupled manipulation	Robust refinement	Sampling event
Robot base mass (kg)	$[-0.5, 0.5]$	$[-1.0, 1.0]$	$[-1.5, 1.5]$	episode reset
Robot friction	$[0.75, 1.20]$	$[0.60, 1.30]$	$[0.50, 1.40]$	episode reset
Motor-strength scale	$[0.95, 1.05]$	$[0.90, 1.10]$	$[0.88, 1.12]$	episode reset
Robot COM offset (m)	$[-0.002, 0.002]$	$[-0.005, 0.005]$	$[-0.010, 0.010]$	simulator start
Joint-position noise (rad)	$[-0.004, 0.004]$	$[-0.008, 0.008]$	$[-0.010, 0.010]$	episode reset
Object-mass scale	disabled	$[0.90, 1.10]$	$[0.80, 1.20]$	episode reset
Object friction	disabled	$[0.70, 1.20]$	$[0.60, 1.30]$	episode reset
Object COM offset (m)	disabled	$[-0.003, 0.003]$	$[-0.005, 0.005]$	simulator start
Horizontal root push ($\mathrm{m\,s^{-1}}$)	disabled	disabled	$[-0.10, 0.10]$ per axis	every 5 s

Coordination-ready approach. The first stage learns stable approach, relative placement, whole-body motion tracking, and controlled stopping around the physical object. Only the approach portion of each paired reference is used. Its endpoint is detected from the first sustained reference-object activity: linear speed at least $0.12\,\mathrm{m\,s^{-1}}$, angular speed at least $0.12\,\mathrm{rad\,s^{-1}}$, or vertical displacement at least $0.04\,\mathrm{m}$ for five consecutive frames. A ten-frame margin is subtracted, and the endpoint is capped at 45% of the sequence; if no sustained activity is detected, 30% of the sequence is used. Sequences with fewer than 24 valid approach frames are excluded. Seventy per cent of episodes start from the first frame, while the remaining 30% start from a uniformly sampled recovery state within 5–85% of the approach segment. Motion-tracking, stability, and inactive-object terms are active, whereas manipulation rewards are disabled. Light robot randomization is used to avoid destabilizing initial skill acquisition.

Coupled object manipulation. The second stage is initialized from the approach policy and expands rollouts to complete approach–lift–carry–place references. The probability of a first-frame reset is annealed from 0.35 to 0.20, and the probability of an early reset within 2–30% of the reference is annealed from 0.30 to 0.20. The remaining probability samples uniformly from 30–95% of the reference; after annealing, the start/early/mid-reference mixture is 0.20/0.20/0.60. The full gated cooperative objective is activated, and randomization is extended to object mass, friction, and centre of mass. This stage establishes complete object manipulation before stronger posture constraints and disturbances are introduced.

Robust posture refinement. The final stage continues from the coupled-manipulation policy while retaining the complete cooperative objective. Arm and wrist posture tracking, stronger lower-body and smoothness regularization, broader robot/object randomization, and periodic horizontal pushes are added. Because the manipulation skill is already established, this continuation improves posture and robustness without requiring the policy to discover contact-coupled behavior under the widest uncertainty from the outset.

Table 11: Teacher ablation design. The table specifies the controlled change and intended comparison only; all variants are evaluated with the common evaluation system described separately.

Variant	Imitation	Object task	Coupling	Gates	Staging/continuation	Purpose
Imitation-only objective	yes	no	no	no	end-to-end	test whether paired tracking alone is sufficient
Decoupled object-manipulation objective	yes	yes	no	no	end-to-end	isolate object-level supervision
Ungated cooperative objective	yes	yes	yes	no	end-to-end	isolate explicit bilateral coupling
Gated cooperative objective (end-to-end)	yes	yes	yes	yes	end-to-end	isolate readiness gating
Gated cooperative objective (stage-wise)	yes	yes	yes	yes	ordered stages	test the full training curriculum
No-randomization continuation	yes	yes	yes	yes	common checkpoint	control for continuation without DR
Full-randomization continuation	yes	yes	yes	yes	common checkpoint	test robustness-oriented DR continuation

All stages use the same policy, value, and discriminator architecture and the same AMP formulation; only the learning rate and entropy regularization are reduced during later continuation. Each stage starts from a checkpoint selected from the preceding stage using a predeclared training budget or validation rule, never test-set performance. The final Teacher checkpoint is fixed before Student training and is shared by all Student variants.

C.3 Controlled Teacher Ablations

The Teacher ablations are organized as three controlled comparisons, as summarized in Table 11. They test (i) which objective components are required to transform paired motion tracking into cooperative object manipulation, (ii) whether readiness gating and physically ordered training improve acquisition, and (iii) whether broad randomization is beneficial when introduced only after the nominal skill has formed. This subsection defines the intervention applied by each variant; all outcomes are measured using the common protocol in Appendix B.

Objective composition and readiness gating. Four end-to-end variants are trained from random initialization using the same full-reference rollout support, reset distribution, nominal randomization, network architecture, optimizer, and training budget. The imitation-only variant retains paired dual-robot motion tracking but disables every object-level and bilateral-coordination term. The decoupled object-manipulation variant adds object lift, tracking, and drop prevention but contains no explicit cross-robot coordination terms. The ungated cooperative variant additionally activates bilateral anchoring and synchronization, while setting readiness gates to one so that all manipulation rewards are immediately available. The gated cooperative variant uses the same reward terms but activates them through Eqs. (42)–(44). The sequential contrasts therefore test object-level supervision, explicit cooperative coupling, and conditional reward activation, respectively.

Stage-wise versus end-to-end optimization. To isolate the optimization schedule from the final reward definition, the end-to-end and stage-wise variants use the same gated cooperative objective and the same total training budget. End-to-end training exposes the policy to the full reference, full gated objective, and target randomization from random initialization. Stage-wise training follows the rollout, reset, reward, and randomization expansion in Table 9. This comparison tests the complete physically ordered curriculum, rather than comparing two different final objectives.

Randomization continuation. The robustness comparison begins from the same selected coupled-object-manipulation checkpoint. Two continuations use identical optimization settings, random seeds, training budgets, and checkpoint-selection rules. The no-randomization continuation disables parameter randomization and external pushes, whereas the full-randomization continuation uses the robust refinement ranges in Table 10. Starting from one common checkpoint prevents differences in initial skill acquisition from being attributed to domain randomization.

All objective-composition variants use matched data, initialization protocol, and compute. The two continuation variants inherit exactly the same policy and optimizer state before their controlled divergence. Multiple independent training seeds are used for every comparison, and model selection is performed without access to the common test set. These controls ensure that observed differences can be attributed to the intended reward, gating, curriculum, or randomization intervention rather than to different data, compute, or Teacher initialization.

Table 12: Proprioceptive observation of each local Student actor.

Component	Dim.
Base angular velocity	3
Projected gravity	3
Relative joint position	29
Relative joint velocity	29
Previous executed normalized action	29
Total	93

Table 13: Object- and task-related observation of each local Student actor.

Component	Dim.
Current object position	3
Current object orientation (6D)	6
Final reference-object goal position	3
Ten reference-object waypoints	90
Total	102

D Local Student-Model Distillation and Deployment

This appendix provides implementation details of the local Student policies, including the local observation interface, the shared-base and role-specific residual architecture, Teacher-success distillation, learner-state refinement, and the online multi-agent reinforcement-learning refinement setup.

D.1 Student Observation Interface

For robot i, the local Student actor receives a 195-dimensional current observation,

$$o_t^i \in \mathbb{R}^{195}. \tag{45}$$

The centralized environment constructs a 262-dimensional canonical observation for each role, including a 67-dimensional robot-reference tracking block. The local Student removes this reference-tracking block and retains only a 93-dimensional proprioceptive block and a 102-dimensional object/task block:

$$195 = 93 + 102. \tag{46}$$

The proprioceptive observation is summarized in Table 12.

Joint positions and velocities are expressed relative to the default robot configuration,

$$q_t^{\mathrm{rel}} = q_t - q_{\mathrm{default}}, \qquad \dot{q}_t^{\mathrm{rel}} = \dot{q}_t - \dot{q}_{\mathrm{default}}. \tag{47}$$

The object/task observation is summarized in Table 13.

The current object state, final object goal, and future object waypoints are all expressed in the current pelvis frame of the corresponding robot role. Object orientation is represented using the first two columns of the rotation matrix, flattened as

$$[m_{00}, m_{01}, m_{10}, m_{11}, m_{20}, m_{21}]. \tag{48}$$

The future object plan consists of ten consecutive reference poses. The reference index of waypoint k is

$$j_k = \min\left(t_{\mathrm{ref}} + k,\, T_{\mathrm{ref}} - 1\right), \qquad k = 0, \ldots, 9, \tag{49}$$

where t_{ref} denotes the current reference frame and T_{ref} is the reference-motion length. Therefore, the waypoint sequence starts from the current reference frame and extends nine frames into the future; when the end of the reference motion is reached, the final frame is repeated.

Object history. In addition to the current observation, each role maintains a 72-dimensional short-term object history,

$$h_t^{o,i} \in \mathbb{R}^{72}, \qquad 72 = 8 \times 9. \tag{50}$$

Each history element consists of a three-dimensional object position and a 6D object orientation,

$$h_{t-k}^{o,i} = \left[p_{t-k}^{o,i}, R_{t-k}^{o,i,6D} \right] \in \mathbb{R}^9. \tag{51}$$

The history buffer internally stores object poses in the world frame. At every policy query, all eight stored poses are re-expressed in the *current* pelvis frame of the corresponding robot and flattened from oldest to newest. The history does not explicitly include object linear or angular velocity.

At episode reset, the post-reset object pose is repeated across all eight history slots rather than zero-padded. The history advances exactly once after each physical environment control transition; querying the actor alone does not modify the history.

Observation normalization. The current observation and object history use separate frozen coordinate-wise normalizers,

$$\hat{x}_j = \frac{x_j - \mu_j}{\sqrt{\sigma_j^2 + 10^{-5}}}. \tag{52}$$

Normalization statistics are computed only from the frozen Teacher-success replay, with equal weighting across motions. The normalizers remain fixed throughout subsequent Student training and deployment.

The production Student does not use an explicit role token. Role information is not concatenated to the actor observation as a scalar or one-hot vector. Instead, role identity is represented by selecting one of two independent role-specific residual branches described below.

D.2 Shared Base and Role-Specific Residual Architecture

The two robots share a substantial amount of underlying control structure, including whole-body balance, support of the common object, and progression along the object-level plan. At the same time, the two grasp sides exhibit systematic asymmetries in body configuration, local contacts, and load distribution. We therefore factorize each local policy into a shared cooperative base and a lightweight role-specific residual:

$$\pi^i = \pi_{\text{shared}} + \Delta\pi^i. \tag{53}$$

Shared cooperative base. The shared base receives the normalized 195-dimensional current observation,

$$z_t^i = B_{\theta_b}\left(\hat{o}_t^i\right), \tag{54}$$

and follows the architecture

$$195 \to 512 \to 256 \to 128 \to 29. \tag{55}$$

All hidden layers use ELU activations. The complete base network is shared between the two roles and outputs a 29-dimensional *unsquashed* base-action mean z_t^i.

Role-specific residual branch. Each role has an independent residual module with architecture

$$206 \to 64 \to 29, \tag{56}$$

again using an ELU hidden activation. To make the initial policy rely primarily on the shared cooperative base, the output-layer weights and biases of both residual branches are initialized to zero.

The residual input is

$$x_t^{r,i} = \left[\hat{o}_{t,0:105}^i, \hat{h}_t^{o,i}, \mathrm{sg}(z_t^i) \right], \tag{57}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Its dimensionality is therefore

$$105 + 72 + 29 = 206. \tag{58}$$

Table 14: Local observation block used by the role-specific residual branch.

Component	Dim.
Base angular velocity	3
Projected gravity	3
Relative joint position	29
Relative joint velocity	29
Previous executed action	29
Current object position	3
Current object orientation (6D)	6
Final object goal position	3
Total	105

The 105-dimensional local feature vector is not a separately designed observation. It is exactly the first 105 dimensions of the normalized 195-dimensional actor observation:

$$105 = 3 + 3 + 29 + 29 + 29 + 3 + 6 + 3. \tag{59}$$

Its composition is summarized in Table 14.

Consequently, the role-specific residual does not directly receive the 90-dimensional future object-waypoint block. Instead, it additionally receives the 72-dimensional object history, allowing the role-specific correction to depend on short-term object motion. The object history is used only by the residual branch and is not provided to the shared base.

The 29-dimensional shared component received by the residual branch is $\mathrm{sg}(z_t^i)$, i.e., the detached base mean *before* the final $\tanh$, rather than the final executed action.

Let

$$u_t^i = R_{\theta_r^i}\left(x_t^{r,i}\right), \tag{60}$$

and define the bounded role residual as

$$r_t^i = \tanh(u_t^i). \tag{61}$$

The shared and role-specific components are combined in latent-action space:

$$\mu_t^i = z_t^i + \lambda_r r_t^i, \tag{62}$$

followed by

$$a_t^i = \tanh(\mu_t^i). \tag{63}$$

The environment boundary additionally applies numerical sanitization and clips the normalized action to $[-1, 1]$.

Student actions are converted into joint-position targets using the reference-centered mapping in Eq. (40). Robot-motion references are therefore used by the action decoder but are excluded from the Student actor input.

D.3 Teacher-Success Distillation

We construct a frozen Teacher-success replay $\mathcal{D}_T$ from successful cooperative rollouts of a frozen centralized Teacher. The replay provides local Student observations and the corresponding role-specific Teacher actions for distillation.

Distillation objective. Let $a_{T,t}^i$ denote the Teacher action for role i. The primary final-action imitation objective is

$$\mathcal{L}_{\mathrm{act}} = \mathbb{E}\left[\left\|\tanh\left(z_t + \lambda_r r_t\right) - a_{T,t}\right\|_2^2\right]. \tag{64}$$

We additionally use a *role-residual proxy loss*, which directly trains the role-specific residual to approximate the Teacher action around a detached shared-base prediction:

$$\mathcal{L}_{\mathrm{role}} = \mathbb{E}\left[\left\|\tanh\left(\mathrm{sg}(z_t) + 0.25\, r_t\right) - a_{T,t}\right\|_2^2\right]. \tag{65}$$

The shared base is stop-gradient in this auxiliary objective, so $\mathcal{L}_{\text{role}}$ updates only the role-specific residual modules. The factor 0.25 is an internal gain used only in this auxiliary proxy and is distinct from the residual gain used in the final Student action.

To prevent the role-specific residual from unnecessarily dominating the shared solution, we also regularize its bounded output:

$$\mathcal{L}_{\text{res}} = \mathbb{E}\left[\|r_t\|_2^2\right]. \tag{66}$$

The complete production distillation objective is

$$\boxed{\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{act}} + \mathcal{L}_{\text{role}} + 10^{-3}\mathcal{L}_{\text{res}}}. \tag{67}$$

$\mathcal{L}_{\text{act}}$ updates both the shared base and the role-specific residuals, whereas $\mathcal{L}_{\text{role}}$ and $\mathcal{L}_{\text{res}}$ only update the residual branches.

Phase-balanced replay. Long-horizon cooperative transport contains semantic phases with substantially different durations. In particular, quasi-steady transport may occupy many more frames than short but contact-critical transitions such as grasping and placement. Uniform frame sampling would therefore over-represent long transport segments.

We partition each successful trajectory into six semantic phases,

$$\{\text{approach}, \text{grasp}, \text{lift}, \text{carry}, \text{place}, \text{release}\}, \tag{68}$$

using motion-specific semantic reference-frame intervals rather than equal-length temporal bins.

Training samples are drawn with replacement using the following hierarchy:

1. uniformly sample one motion;
2. uniformly sample one successful Teacher episode for the selected motion;
3. uniformly sample one non-empty semantic phase;
4. uniformly sample one replay row within that phase.

In addition, one episode-frame-zero sample is forced every 64 draws. Since every frozen replay episode contains all six phases and frame zero belongs to the approach phase, the resulting design probability for approach is

$$P(\text{approach}) = \frac{1}{64} + \frac{63}{64}\frac{1}{6} = \frac{23}{128}, \tag{69}$$

whereas each of the other five phases has probability

$$P(p) = \frac{63}{64}\frac{1}{6} = \frac{21}{128}. \tag{70}$$

This produces an approximately phase-balanced training distribution while retaining a small number of frame-zero anchor samples.

Nominal optimization. The nominal Student is trained using

$$30{,}000{,}128 \tag{71}$$

sampled replay draws. Each outer iteration contains 256 rows, divided into four 64-row minibatches and reused for five epochs. This corresponds to 20 Adam optimizer steps per outer and

$$2{,}343{,}760 \tag{72}$$

optimizer updates over the complete nominal-distillation stage.

The learning rate is

$$2.5 \times 10^{-5} \tag{73}$$

for the shared base and

$$10^{-4} \tag{74}$$

for each role-specific residual branch. We use Adam with zero weight decay and gradient-norm clipping at 1.0. Observation-normalization statistics remain frozen throughout this stage.

D.4 Learner-State Refinement

Teacher-success distillation primarily covers states visited under successful Teacher execution. During deployment, however, the Student itself drives the closed-loop dynamics. Consequently, its state distribution generally differs from that of the Teacher:

$$d^{\pi_S}(s) \neq d^{\pi_T}(s). \tag{75}$$

Even when the Student achieves low imitation error on Teacher-success states, small action errors can accumulate through closed-loop execution, changing robot posture, contact geometry, and shared-object motion. Because the two robots are physically coupled through the shared object, these deviations also change the future interaction state observed by the partner.

To reduce this distribution shift, we perform learner-state refinement after nominal distillation. The current Student is first rolled out in closed loop,

$$s_t \sim d^{\pi_S}, \tag{76}$$

while the centralized Teacher performs a privileged shadow query at the visited Student-induced state,

$$a_{T,t} = \pi_T\left(o_t^{\text{priv}}\right). \tag{77}$$

This produces a learner-state replay

$$D_L = \left\{(o_t^i, a_{T,t}^i) \mid s_t \sim d^{\pi_S}\right\}. \tag{78}$$

The learner-state replay is then mixed with the original Teacher-success replay:

$$D_{\text{mix}} = (1-\rho)D_T + \rho D_L, \tag{79}$$

and the Student is updated through a short re-distillation stage.

D_T preserves the nominal cooperative behavior demonstrated by successful Teacher trajectories, whereas D_L expands training coverage toward states that the Student actually encounters during its own closed-loop execution. The mixing coefficient ρ controls the amount of learner-induced supervision while retaining Teacher-success replay as the primary training distribution.

D.5 Online Cooperative RL Refinement

After learner-state refinement, we perform a final cooperative reinforcement-learning refinement using MAPPO. The refined Student policy is kept frozen, and MAPPO optimizes only an additional residual actor together with a centralized critic. This stage refines closed-loop cooperative behavior without modifying the shared base or the internal role-specific residual branches learned during Student distillation.

The external residual actor is shared between the two roles and receives an explicit 2D role one-hot vector. Its input therefore consists of the 195-dimensional local observation, the 72-dimensional object history, and the 2-dimensional role encoding:

$$269 = 195 + 72 + 2. \tag{80}$$

The residual actor follows

$$269 \to 256 \to 128 \to 29, \tag{81}$$

with ELU hidden activations and a zero-initialized output head.

The centralized critic receives the full 524-dimensional team observation and uses the architecture

$$524 \to 1024 \to 512 \to 256 \to 128 \to 1. \tag{82}$$

The role representation therefore differs between the distilled Student and the external MAPPO residual: the former represents role asymmetry through two independent role-specific residual branches, whereas the latter uses a shared residual actor conditioned on an explicit role one-hot vector.

Importantly, the external MAPPO residual is distinct from the two role-specific residual branches of the distilled Student. During this refinement stage, the shared base and both internal role-specific residual branches remain frozen. Only the external residual actor and the centralized critic are optimized by MAPPO.

Residual policy. Let $a^i_{t,S}$ denote the action produced by the frozen Student after learner-state refinement. The external MAPPO residual actor predicts a Gaussian latent action,

$$z^i_t \sim \mathcal{N}\left(\mu^i_{\mathrm{ext},t}, 0.2^2 I\right). \tag{83}$$

During online training, the sampled residual is added to the frozen Student action:

$$a^i_{t,\mathrm{RL}} = \mathrm{clip}\left(a^i_{t,\mathrm{S}} + 0.1\tanh(z^i_t), -1, 1\right), \tag{84}$$

where $a^i_{t,\mathrm{S}}$ denotes the frozen Student action and the scalar 0.1 controls the authority of the external residual.

For deterministic evaluation, exploration noise is removed and the residual distribution mean is used:

$$a^i_{t,\mathrm{RL}} = \mathrm{clip}\left(a^i_{t,\mathrm{S}} + 0.1\tanh\left(\mu^i_{\mathrm{ext},t}\right), -1, 1\right). \tag{85}$$

This residual parameterization restricts online reinforcement learning to local corrections of the existing Student behavior rather than directly re-optimizing the complete Student network.

MAPPO optimization. Each MAPPO refinement rollout uses 70 parallel environments and 64 control steps, yielding

$$70 \times 64 = 4{,}480$$

team transitions per update. A team transition corresponds to one environment control step in which both robots act jointly.

We use

$$\gamma = 0.99, \qquad \lambda_{\mathrm{GAE}} = 0.95,$$

together with a PPO clipping coefficient

$$\epsilon_{\mathrm{PPO}} = 0.05.$$

Each rollout is followed by a single full-batch PPO update (`ppo_epoch`=1). The actor and critic learning rates are 10^{-4} and 3×10^{-4}, respectively. The centralized critic uses running ValueNorm and a 64-step target/GAE horizon.

Each MAPPO refinement run consists of four rollout–update cycles, corresponding to

$$4 \times 4{,}480 = 17{,}920$$

additional team transitions after Student refinement. All MAPPO results reported under this refinement protocol use the final checkpoint after the fixed 17,920-transition budget. Intermediate evaluation scores are not used for checkpoint selection. The 17,920 transitions therefore refer only to the MAPPO refinement stage and do not include Teacher training or Student distillation.

Optimization ablations. The BC-only variant uses the same refined Student checkpoint but disables the external reinforcement-learning residual, so that the Student action is executed directly without MAPPO refinement. The IPPO variant uses the same external residual actor initialization, reward, interaction budget, and checkpoint rule as MAPPO, while replacing the centralized critic with an independent critic. Learner-state refinement is kept unchanged in these optimization ablations and is evaluated separately in the corresponding distillation ablation.