DUET: Dual-Robot Understanding via Efficient Teaching

Yiqi Zhao[1]*, **Ruohai Ge**[1]*, **Celina Shiyu Wang**[1], **Junjie Ye**[1], **Muchen Xu**[1], **Minhao Li**[1], **Sergey Zakharov**[2], **Basile Van Hoorick**[2], **Vitor Campagnolo Guizilini**[2], **Leonidas Guibas**[3], **Gaurav S. Sukhatme**[1], **Jyotirmoy V. Deshmukh**[1], **Yue Wang**[1]

[1]University of Southern California [2]Toyota Research Institute [3]Stanford University

Figure 1: **DUET.** We introduce a dual-robot policy learning framework that features efficient learning from human demonstrations. Left: Two humans perform a collaborative manipulation task. Right: A heterogeneous dual-robot system mimics the human teachers.

Abstract: Dual-robot collaboration enables tasks that exceed the reach and payload of a single robot, such as collaboratively transporting objects across environments and executing coordinated handovers. Data acquisition is the primary bottleneck for training these systems. To this end, we introduce DUET, a dual-robot learning framework for mobile manipulation. For efficient data collection, we create a unified dual-embodiment synchronized VR-based teleoperation system for in-domain heterogeneous robot data collection. We further develop a complementary tracking pipeline that records human-human coordination and collaborative mobile manipulation priors. To allow efficient learning, we introduce an Action Chunking Transformer based architecture that first pretrains collaborative policies on efficient human-human demonstrations, before finetuning them on a minimal set of real-robot teleoperation trajectories. We develop a benchmark of four collaborative tasks to evaluate our framework using a Unitree G1 humanoid and a Dexmate Vega1 mobile manipulator. The results demonstrate that harnessing human priors not only yields superior task performance compared to baselines trained only on robot data, but also reduces the total human effort required for data collection. Our human data collection pipeline achieves $5.4\times$ acceleration on average from teleoperation, but we perform equally or better than robot-only data trained policies across all tasks. Our project page is available at https://zhaoy37.github.io/Duet/.

Keywords: Dual-robot Collaboration, Learning from Human Demonstrations

*Equal contribution. Order is determined by coin flip.

1 Introduction

Many real-world tasks intrinsically exceed the kinematic reach and payload limits of a single robot [1]. Operations that require simultaneous, spatially separated interactions, such as handling oversized objects or executing asymmetric tool-use, make dual-robot systems a physical necessity rather than merely an algorithmic extension [2, 3]. Yet, realizing their practical utility requires advancing beyond collision-free navigation toward *contact-rich collaborative mobile manipulation* [4]. Controlling heterogeneous robot duos in this regime is not a trivial superposition of independent policies; it demands tightly coupled spatial-temporal coordination. For example, jointly manipulating cumbersome shared payloads demands strict dynamic equilibrium, while asymmetric interactions require tight temporal synchronization and shared perception. Because analytically modeling the multi-body dynamics of such joint operations is computationally prohibitive and brittle, end-to-end visuomotor policies [5, 6] have emerged as the compelling alternative, directly mapping high-dimensional observations to coordinated actions while bypassing explicit state estimation.

However, realizing the potential of these policies is severely bottlenecked by data acquisition. Orchestrating demonstrations for heterogeneous dual-robot systems typically demands complex setups and a large pool of human operators. To address this, we design a synchronized dual-operator teleoperation framework that allows just two human experts to successfully command the heterogeneous duo, a significant step forward in dual-robot data collection. Yet, even with this efficient system, teleoperation remains inherently limited by time and cost. To scale our dataset, we complement this framework with a highly efficient human-human data collection pipeline where two humans directly execute the collaborative tasks. Capturing direct human demonstrations [7, 8, 9] is fundamentally faster and completely decouples data volume from robot hardware constraints, while naturally encoding rich, transferable priors regarding multi-robot spatial coordination. To synthesize these complementary streams, we introduce a pretraining paradigm that leverages low-cost human demonstrations to bootstrap the policy, requiring only a minimal set of high-fidelity dual-operator robot data. This approach effectively bridges the human-to-robot gap, maximizing the utility of both human agility and our streamlined dual-teleoperation system.

We introduce DUET (**D**ual-robot **U**nderstanding via **E**fficient **T**eaching), a unified centralized visuomotor policy framework for heterogeneous dual-robot collaboration (Figure 1). While supported by a dedicated dual-robot teleoperation pipeline, DUET fundamentally bypasses scaling limits by integrating efficient human demonstrations harnessing SAM 3D Body [10]. To bridge the resulting cross-embodiment gap, all data is projected into a shared pose space. Conditioned on these unified states and egocentric RGB streams, a single Action Chunking Transformer (ACT) [11] backbone intrinsically enforces the tight spatial-temporal coupling demanded by contact-rich tasks. We validate DUET on a Unitree G1 humanoid (G1) and a Dexmate Vega1 mobile robot (Vega1) across a novel benchmark of four contact-rich tasks. Empirically, DUET demonstrates that integrating these human priors via collective training achieves the same or better task performance compared to fully robot-demonstrated baselines, while requiring less overhead from teleoperation. We summarize our contributions:

1. **Heterogeneous Dual-Operator Teleoperation:** We design a two-operator teleoperation framework to collect in-domain trajectories for a heterogeneous duo.
2. **Human Collaboration Pipeline:** We propose a scalable data collection pipeline that captures direct human-human collaboration, ready for cross-embodiment pretraining.
3. **Collective Training Architecture:** We harness an ACT-based framework that leverages human and robot data streams for collective training.
4. **Benchmark and Evaluation:** We introduce a novel benchmark of four tasks including sweeping, transferring, balancing, and handovers. Empirically, pretraining on human demonstrations lowers overall data collection effort while achieving higher success rates than policies trained solely on larger robot datasets.

2 Related Work

Imitation Learning and Teleoperation. Imitation learning via human teleoperation has emerged as one of the premier paradigms for scaling physical interaction in open-world environments [11, 12, 13]. Recent breakthroughs have expanded this paradigm by integrating manipulation with mobile bases and legged locomotion, resulting in robust teleoperation pipelines for wheeled bimanual platforms [4, 14, 15], scalable in-the-wild data collection interfaces [16, 17, 18], and legged loco-manipulation frameworks for humanoids [19, 20, 21, 22, 23]. However, these infrastructures are designed for single-robot, leaving high-fidelity data collection for spatially distributed, heterogeneous multi-robot teams unresolved [24, 25]. To bridge this gap, we introduce a portable, VR-based teleoperation framework for egocentric dual-robot data collection in mobile manipulation tasks.

Multi-Robot Collaboration. Multi-robot physical collaboration has a rich history, traditionally relying on classical control paradigms that demand complex state estimation, rigorous system modeling, and explicit communication architectures [26, 27, 28, 29]. Seeking more adaptable behaviors, recent advances have increasingly shifted toward learning-based methods [30, 31]. Reinforcement learning can elicit sophisticated cooperative strategies in simulation, but typically depends on privileged, low-dimensional states and struggles to generalize to unstructured, visually rich real-world environments [32, 33, 34, 35, 36]. Concurrently, end-to-end vision-action models have demonstrated remarkable generalizability for real-world manipulation tasks [37, 12, 38]. To overcome the generalization limits of state-dependent control, we introduce a heterogeneous dual-robot framework directly mapping shared visual observations to synchronous joint actions.

Learning from Human Demonstrations. To alleviate the slow and costly nature of on-robot data collection, human demonstrations offer a scalable alternative [9]. Capturing rich physical affordances, human data provides a structural prior for policy learning. Recent frameworks leverage these priors via visual representation pre-training or cross-domain co-training to successfully transfer human dexterity to robots [8, 7, 39]. This paradigm has rapidly advanced complex embodiments, enabling view-invariant skill transfer [40], in-the-wild humanoid loco-manipulation [19, 41, 42], and interactive whole-body control via human demonstrations [43, 44, 45]. However, while these priors have scaled single-robot learning, existing multi-robot extensions primarily target social or unconstrained kinematic behaviors. Utilizing synchronized multi-human data for physically coupled, object-centric manipulation remains unexplored. We introduce a multi-human pretraining pipeline, leveraging collaborative human data as a prior to accelerate and enhance our visuomotor policy.

3 Method

3.1 Teleoperation Pipeline

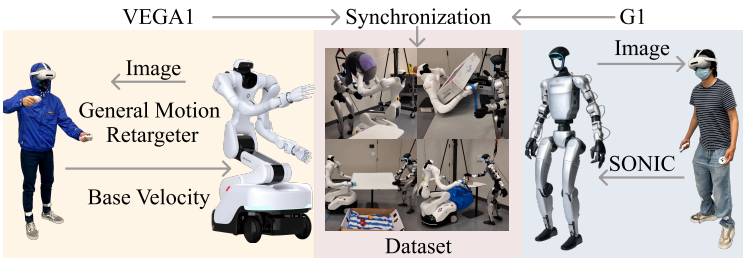


Figure 2: **Dual-Robot Teleoperation Pipeline.** Two human operators utilize PICO VR interfaces to simultaneously control a heterogeneous robot duo, Vega1 via General Motion Retargeting [46] and G1 via the SONIC [20] framework. The system streams real-time visual feedback while asynchronously logging joint pose data.

As illustrated in Figure 2, we introduce a synchronized teleoperation framework that enables just two human operators to simultaneously command a heterogeneous duo, demonstrated using a G1 and a Vega1. This system addresses the data acquisition bottleneck in multi-robot learning by providing a streamlined, low-latency interface for collaborative loco-manipulation tasks. By decoupling the human interface from the underlying execution layers, the architecture serves as a mature, hardware-

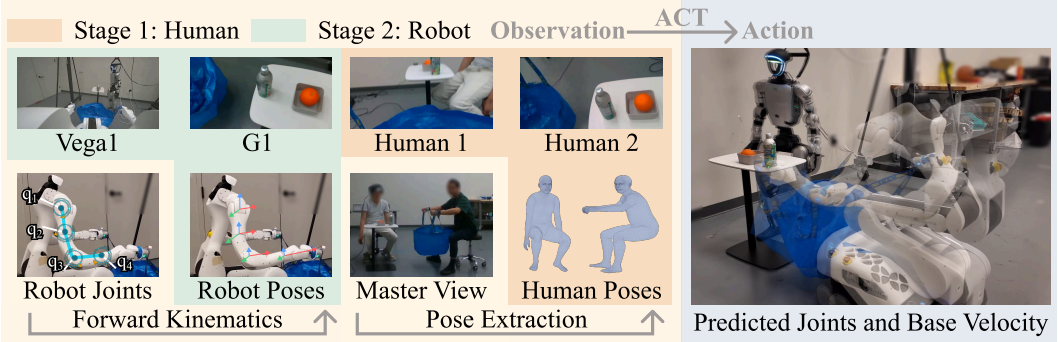


Figure 3: **DUET Overview.** In *stage* 1, we pretrain an ACT architecture (Figure 4) with RGB streams from human-human demonstrations and extracted keypoints for proprioceptives. In *stage 2*, in-distribution robot data finetunes the ACT policy.

agnostic foundation. It is capable of supporting any robotic embodiment with a compatible action retargeting module, allowing new platforms to be hot-swapped into the collaborative pipeline.

Hardware and Visual Interface: To ensure high-fidelity spatial awareness during physical interactions, operators utilize PICO VR headsets paired with handheld controllers and two ankle trackers. This setup serves as a comprehensive motion-tracking interface while simultaneously handling real-time egocentric video streaming via GStreamer. The captured human kinematic data is continuously streamed to the robot execution layers, while the visual feedback is explicitly tailored to each platform's sensory suite: the G1 operator receives a 1920×1080 (30 Hz) stream from a RealSense camera, while the Vega1 operator receives a wide-angle 2560×720 (30 Hz) stream from a ZED camera. To eliminate the standard requirement of a dedicated third operator to manage host PC interventions during rapid data collection, the VR controllers also serve as the central command hub, directly managing task initialization, environment resets, and data logging.

Modular Tracking and Retargeting: Mapping strategies are tailored to the physical capabilities of each platform. For G1, we utilize the state-of-the-art SONIC [20] framework for physics-based, whole-body motion tracking. However, to adapt this for collaborative teleoperation, we introduce several critical extensions: we implemented the aforementioned real-time image streaming and integrated a 10 Hz asynchronous data collection module. Conversely, the control pipeline for Vega1 is entirely custom-designed, with the exception of the General Motion Retargeting (GMR) [46] algorithm used for retargeting human motion to robot action. We designed a GMR configuration that explicitly retargets upper-body human motion to Vega1 and mapped the PICO controller's analog stick to mobile base's v_x, v_y and ω_z. We implemented a lower-level inference controller utilizing Exponential Moving Average smoothing for stability, alongside visual and data logging modules similar to G1. Finally, to synchronize our collected dual-robot data, both robots log sensor and action states asynchronously at 10 Hz against host clocks synchronized via a local Chrony NTP server, enabling exact automated alignment during offline post-processing.

3.2 Human Data Collection Pipeline

To complement the teleoperation system, we develop a pipeline that records two humans performing the same collaborative tasks without a robot in the loop. The recording setup consists of three cameras. We use one head/neck-mounted camera for each operator and a fixed third-view *master* camera positioned to keep both operators in frame, providing the input to subsequent offline human pose extraction following a soft cross-camera time alignment.

For each human clip, we start by extracting the poses of the two operators. We harness a YOLO-based [47] human detector to produce a set of 2 bounding boxes tracking the human identities. For each human, SAM 3D Body [10] runs per frame for mesh extraction that lives in an arbitrary pseudo-camera frame with an unknown global scale. To lift the predictions into metric coordinates, we fuse the SAM predicted mesh with the master camera's depth stream. For each frame and

tracked person, we render the mesh through SAM's internal focal length to obtain a predicted z-buffer $\hat{z}(u, v)$, restrict comparison to an eroded person silhouette, and recover the metric scale s from the per-pixel ratio between the observed and predicted depth with a robust two-stage estimator $s_0 = \text{median}_{(u,v)\in\mathcal{M}} \frac{z^{\text{obs}}(u,v)}{\hat{z}(u,v)}$, $s = \frac{\sum_{(u,v)\in\mathcal{I}} z^{\text{obs}}(u,v)\,\hat{z}(u,v)}{\sum_{(u,v)\in\mathcal{I}} \hat{z}(u,v)^2}$, where $\mathcal{M}$ is the eroded silhouette and $\mathcal{I} \subseteq \mathcal{M}$ contains pixels for which the rescaled prediction $s_0\hat{z}$ agrees with the observation to within a tolerance threshold. We back-project SAM's 2D keypoints through the camera intrinsics at the recovered metric depths, yielding 3D poses in the physical camera frame of the master view. Lastly, a per-keypoint mask-weighted moving-average filter suppresses high-frequency jitter introduced by depth-sensor noise. We extract 9 3D keypoints (27 dimensions) per human and root velocities from pelvis movements, which we then use to pretrain our visuomotor policy.

3.3 Collective Training

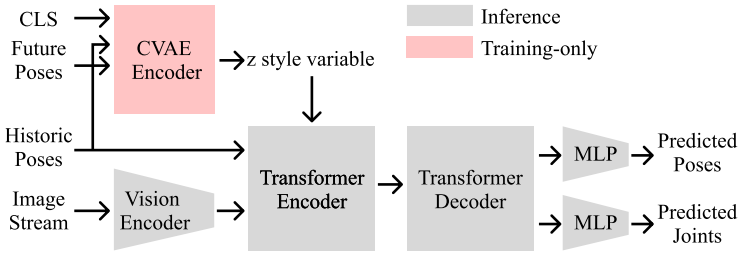


Figure 4: **Model Architecture.** Our ACT policy leverages a 2-head design with the joint head randomly initialized during finetune.

Our robots are controllable via *joint-space* commands. Our architecture (Figure 4) is inspired by [7] and we train a single policy in two stages that share an ACT [11] based architecture but differ in the data source. As depicted in Figure 3, **Stage 1** pretrains the policy backbone on human-human collaboration data collected following Section 3.2; **Stage 2** finetunes the resulting weights on robot teleoperation data of Section 3.1. The two stages produce models with nearly identical parameter shapes, so all components except robot-specific projections are warm-started from the human prior.

We consider an action representation where $[\hat{p}_{t+H}, \hat{q}_{t+H}]$ is the H length predicted next-chunk position with $p_t \in \mathbb{R}^{d_p}$ encoding the *(3D) pose* component shared by both embodiments and $q_t \in \mathbb{R}^{d_q}$ encoding the *joint-space* component available only for robots. We use $\hat{p}$ to denote the prediction of ground-truth p and similarly define $\hat{q}$. To bridge the embodiment gap, both modalities express p_t in head-relative coordinates: Human poses p_t^H recovered by Section 3.2 are re-centered at the operators' heads, while robot poses p_t^R obtained from forward kinematics are expressed relative to each head link. For our heterogeneous duo, $d_p = 57$ comprises 27 dimensions each of G1 and Vega1 keypoints and 3 of Vega1 mobile base velocity command. On the human side, we derive the keypoints of the corresponding poses from Section 3.2 and estimate the base velocity from pelvis.

We utilize an ACT policy with a ResNet-18 [48] visual encoder warm started from ImageNet [49], conditioned on a pair of historic egocentric RGB image streams (one per agent) and the corresponding pose-only proprioceptive states. To accommodate the heterogeneous action space between humans and robots, we adapt the architecture from [7] and project output tokens of the ACT decoder backbone via two parallel multi-layer perceptrons that share the decoder's hidden state $h_t \in \mathbb{R}^{d_{\text{model}}}$ but predict separately $\hat{p}_{t+H} = \text{MLP}_p(h_t)$ and $\hat{q}_{t+H} = \text{MLP}_q(h_t)$. Both heads actively backpropagate on robot batches, while only the pose head is active during human batches. For structured regularization for the action-chunk decoder, our CVAE encoder ingests the pose-only subspace of the action chunk in both stages, ensuring consistent latent semantics during both stages. Adopting the standard [11], during inference, the CVAE encoder is bypassed and the decoder is conditioned on $z = \mathbf{0}$. The joint-space predicted pose and base velocity are forwarded for robot execution.

4 Experimental Result

We evaluate the efficacy of DUET through a series of experiments designed to characterize both our data collection pipelines and the resulting visuomotor policies. Our empirical evaluation aims

Task	Pipeline	Total Time (minute)	Episodes
T1: Trash Collection	Teleoperation	43	86
	Human Demonstration	35.7	107
T2: Box Organization	Teleoperation	71	71
	Human Demonstration	53.3	80
T3: Board Tilting	Teleoperation	22	66
	Human Demonstration	21.5	99
T4: Doll Passing	Teleoperation	31.5	63
	Human Demonstration	26.3	79

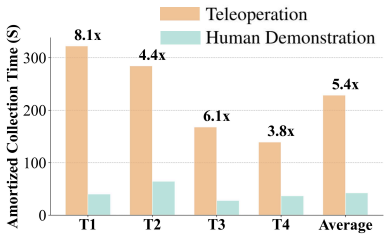


Figure 5: **Benchmark data distribution and collection efficiency. Left:** Total data span and number of episodes per pipeline across four tasks. **Right:** *Amortized collection time* (AT) for each task. The value above each task group is the *speedup ratio*, defined as the teleoperation AT divided by the human-demonstration AT, where AT is the average time for each successful data collection.

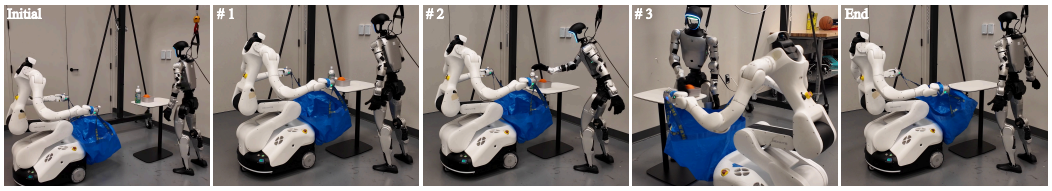
to answer the following research questions: **a) Q1:** How does the collection efficiency of direct human collaboration compare to our streamlined dual-robot teleoperation baseline? **b) Q2**: Does our dual-operator teleoperation pipeline serve as an effective source for high-fidelity dual-robot policy learning? **c) Q3:** Can our human data pretrained policy achieve better/equivalent performance on mobile manipulation tasks while requiring less effort in data collection? We remark that the performance of our framework is dependent on our choice of the ACT architecture, substitutable with alternative visuomotor backbones without changing the data collection and pretraining framework. Our evaluation focuses on demonstrating the feasibility of learning end-to-end dual-robot visuomotor policies, while quantifying the gains achieved by integrating human priors.

4.1 Benchmark and Evaluation Setup

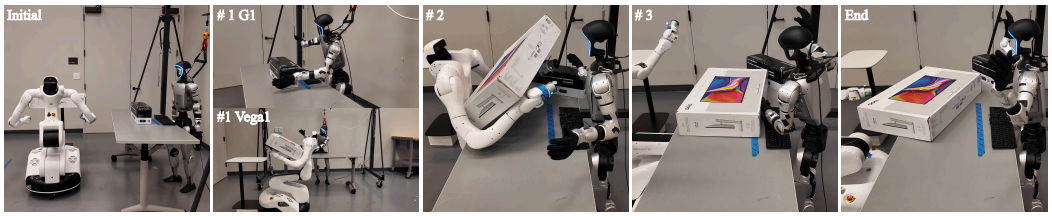
We introduce a benchmark suite of four tasks designed to isolate key multi-robot challenges, as detailed in Figure 6. For each task, we use our teleoperation pipeline (Sec. 3.1) to collect high-fidelity robot data on a heterogeneous platform, comprising Vega1 and G1, alongside corresponding human-human demonstrations (Sec. 3.2): **a) T1: Trash Collection** focuses on *asymmetric spatial-temporal coordination and time synchronization*. Vega1 navigates from a starting location to dynamically position a trash bag as G1 sweeps debris off a table. **b) T2: Box Organization** focuses on *collaborative vision and long-horizon execution ($\sim$ 1 minute)*. G1 clears the workspace. Vega1 transfers a box to its perception boundary, where G1 then executes a corrective nudge. **c) T3: Board Tilting** focuses on *collaborative manipulation and collective balance*. G1 and Vega1 cooperatively tilt a shared whiteboard to slide surface objects into a container, requiring precise force coordination to prevent the board from falling. **d) T4: Doll Passing** focuses on *coordinated handovers and grasp transfer*. Vega1 secures and transfers a doll directly to G1, requiring synchronized release and capture maneuvers. The total number of collected episodes for each task across both teleoperation and human demonstration is detailed in Figure 5; The dataset will be released at https://zhaoy37.github.io/Duet/. To maintain a stable movement speed across the dataset, we account for the inherent kinematic overhead of teleoperation [7]. Because human execution is naturally faster than the robots, the total duration of each human data clip is proportionally reduced by a factor of 1.5 compared to the robot trajectories, and we proportionally scale the velocity in the training data.

4.2 Data Collection Efficiency Comparison

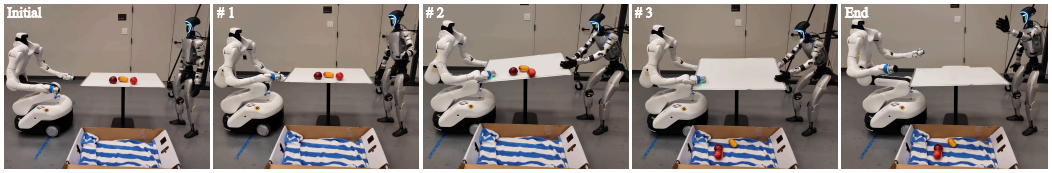
To address **Q1**, we evaluate data acquisition efficiency by comparing the *Amortized Collection Time* (AT) of both pipelines, where AT is the ratio between the total elapsed time devoted for data collection and the number of successfully acquired trajectories ($N = 10$) during this time. As shown in Figure 5 (right), we report an AT of **322.1**, **284.2**, **167.9**, and **139.2** seconds for teleoperation and an AT of **39.9**, **64.4**, **27.7**, and **36.6** seconds for human demonstration respectively for **T1, T2, T3** and **T4**. While our teleoperation framework is highly optimized, with an average AT of **228.4** seconds per task, direct human demonstration provides a **5.4**$\times$ acceleration on average. This underscores the advantages of human data collection: it improves collection efficiency while bypassing the high capital expenses of robot hardware and the severe cognitive bottlenecks inherent to teleoperation.



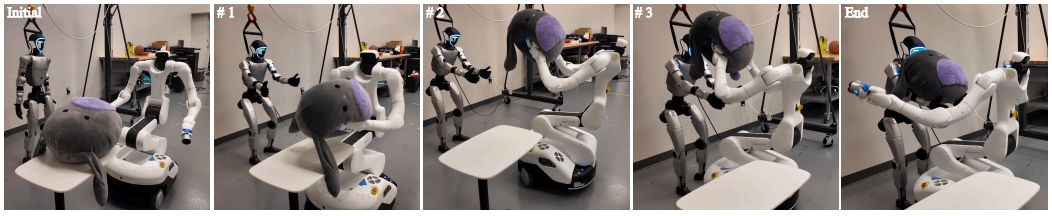
(T1: Trash Collection) Vega1 carries a trash bag and G1 stands by a table with a bottle and a punnet with an orange **(Initial)**. Vega1 approaches the table **(#1)**. G1 pushes the bottle into the trash bag **(#2)**. G1 pushes the punnet with the orange into the trash bag **(#3)**. The trash has been cleaned **(End)**.



(T2: Box Organization) Vega1 stands in front of a white box and G1 stands behind a table with a black box **(Initial)**. Vega1 carries the white box and G1 clears the black box to the side of the table **(#1)**. Vega1 delivers the box onto the cleared table with the box possibly crossing the blue line **(#2)**. G1 pushes the box to prevent it from crossing the blue line **(#3)**. The box has been successfully transported under tight physical and perceptual constraints **(End)**.



(T3: Board Tilting) G1 and Vega1 stand on the sides of a white board with three fruits on top **(Initial)**. G1 and Vega1 approach the white board **(#1)**. G1 and Vega1 collaboratively tilt the white board **(#2)**. The fruits fall into the basket **(#3)**. G1 and Vega1 secure the board on the table with fruits in basket **(End)**.



(T4: Doll Passing) Vega1 stands in front of a doll away from G1 **(Initial)**. Vega1 holds the doll **(#1)**. Vega1 turns and moves toward G1 **(#2)**. Vega1 hands over the doll to G1 **(#3)**. G1 holds the doll **(End)**.

Figure 6: **Overview of the Tasks.** Snapshots show the initial setup and phases.

4.3 Robot-only Ablation Study

To address **Q2**, we conduct an ablation study where only **Stage 2** from Section 3.3, training with 50 robot-only data for each task, is carried out on a fresh ACT (implemented through [50]) without human prior. To quantify multi-stage performance, we implement a normalized metric scoring each task trial from 0.0 to 1.0 based on sequential milestone completion. For **T1**, each piece of trash deposited into the basket yields 0.5 points for a total of two items. For **T2**, 0.5 points are awarded when G1 clears the workspace while Vega1 successfully secures the box, and the remaining 0.5 points are granted if the box is placed on the table without crossing the boundary line, which includes successful recoveries via a corrective nudge from G1. For **T3**, 0.5 points are awarded if the whiteboard remains on the table and at least one of the three items lands in the basket, with the full 1.0 point achieved only when all three items are safely contained. For **T4**, the initial object grasp by Vega1 and the subsequent dual-robot handover each contribute 0.5 points. To evaluate policy performance, we execute **10** independent physical hardware trials per task. As shown in Figure 7, the policy yields overall success rates of **40%, 50%, 60%, 20%** respectively for **T1, T2, T3**, and **T4**. The cumulative scores across all 10 trials are **6.5, 7, 6.5**, and **5.5** respectively for the four tasks,

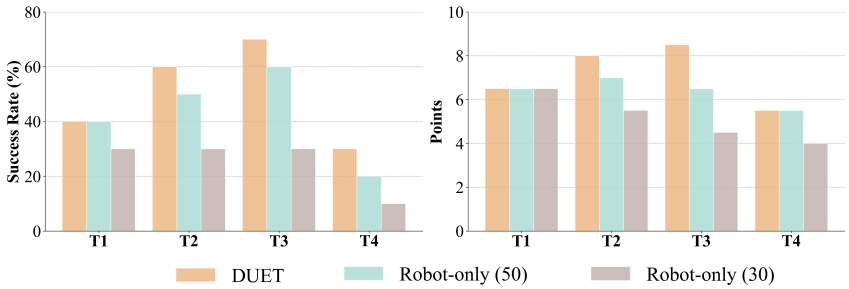


Figure 7: **Evaluation Results.** Comparison of task success rates and accumulated points over 10 experiments for each task of T1–T4. DUET (60 human demos + 30 robot trajectories) is compared against robot-only baselines (trained on 30 and 50 trajectories).

demonstrating that our teleoperation pipeline yields a capable standalone baseline for multi-robot learning.

4.4 DUET Performance

To address **Q3**, we execute **Stage 1** (Section 3.3) by pretraining on 60 human demonstrations, followed by **Stage 2**, where we finetune using 30 robot data. We evaluate these policies and present the resulting success rates and points in Figure 7, where each experiment is conducted 10 times. In this figure, we also compare our policy against the baseline policies trained on 30 and 50 robot only data without human prior. The results demonstrate a clear performance gain or equal performance for the policy pretrained on human data across all evaluated tasks. Drawing on the efficiency metrics established in Section 4.2, we compute the *collection effort* (E) for each policy, measuring the total expected time required to acquire its underlying training dataset. We report these metrics in Table 1. Formally, effort is defined as $E = N_H \cdot AT_H + N_R \cdot AT_R$, where N_H and N_R denote the number of human and robot data used, while AT_H and AT_R represent their respective amortized collection times computed in Section 4.2. We show that while DUET achieves better performance, its training requires less collection effort than policies trained on 50 robot-only data. Qualitatively, the human-pretrained policy exhibits smoother execution and better adheres to object affordance priors implicitly captured during human data collection. These factors largely account for its higher success rate compared to the baseline.

Method	T1	T2	T3	T4
DUET (Ours)	200.95	206.5	111.65	106.2
Robot-only (50)	268.42	236.83	139.92	116
Robot-only (30)	161.05	142.1	83.95	69.6

Table 1: **Data Collection Effort in Minutes.**

5 Conclusion

In this work, we introduced a multi-human pre-training pipeline, leveraging collaborative human data as a prior to accelerate and enhance our dual-robot visuomotor policy. Our evaluations demonstrate that policies pre-trained on human data yield task performance that meets or exceeds baselines trained exclusively on robot data of comparable collection time. Furthermore, we observe that incorporating human data can yield high motion smoothness, attributable to the diverse and smooth nature of human movement. A primary takeaway is that, despite the inherent domain gap, human demonstrations can effectively reduce our reliance on robot data through an appropriate pre-training framework. More broadly, this underscores the immense potential of large-scale human data as a pre-training foundation for multi-robot design. Building on this paradigm, future systems can seamlessly adapt these multi-human priors to diverse embodiments via minimal fine-tuning, paving the way for scalable coordination across any number of robots.

6 Limitation

Model Exploration. While our framework is model-agnostic, we only experimented on the ACT architecture. Further explorations with other state-of-the-art architectures, which may improve the pretraining quality, are left as future work.

Embodiment and Team Size. Our evaluation is restricted to one heterogeneous pair. Our pipeline is designed to be hardware-agnostic, and extending DUET to a wider range of embodiments and larger robot teams is a natural next step toward general-purpose, scalable multi-robot collaboration.

7 Acknowledgments

We thank Cameron Smith for valuable discussions on 3D computer vision and general robot learning. We thank Hongyi Jing and Yuzhe Qin for sharing their expertise and answering our questions about the robots, and Enze Li and Kai Liang for their help with robot repairs. We are also grateful to everyone in the RESL Lab at USC and Yihe Tang for their valuable discussions and feedback on paper writing. The USC Physical Superintelligence Lab acknowledges generous support from Toyota Research Institute, Dolby, Google DeepMind, Capital One, Nvidia, Bosch, NSF, and Qualcomm. This work was partially supported by the National Science Foundation through the following grants: CAREER award (SHF-2048094), IIS-SLES-2417075, and funding by Toyota R&D through the USC Center for Autonomy and AI. Yue Wang is supported by a Powell Research Award.

References

[1] J. Fink, M. A. Hsieh, and V. Kumar. Multi-robot manipulation via caging in environments with obstacles. In *2008 IEEE International Conference on Robotics and Automation*, pages 1471–1476. IEEE, 2008.

[2] M. Lai, K. Go, Z. Li, T. Kröger, S. Schaal, K. Allen, and J. Scholz. Roboballet: Planning for multirobot reaching with graph neural networks and reinforcement learning. *Science Robotics*, 10(106):eads1204, 2025.

[3] Z. Wang and M. Schwager. Multi-robot manipulation without communication. In *Distributed autonomous robotic systems: The 12th international symposium*, pages 135–149. Springer, 2016.

[4] Z. Fu, T. Z. Zhao, and C. Finn. Mobile ALOHA: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=FO6tePGRZj.

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. OpenVLA: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ZMnD6QZAE6.

[6] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, S. Indupuru, Y. L. Tan, C. Zhu, J. Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[7] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025.

[8] R. Hoque, P. Huang, D. J. Yoon, M. sivapurapu, and J. Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FFxkFMU89E.

[9] R. Punamiya, S. Kareer, Z. Liu, J. Citron, R.-Z. Qiu, X. Cai, A. Gavryushin, J. Chen, D. Liconti, L. Y. Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

[10] X. Yang, D. Kukreja, D. Pinkus, A. Sagar, T. Fan, J. Park, S. Shin, J. Cao, J. Liu, N. Ugrinovic, et al. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

[11] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi:10.15607/RSS.2023.XIX.016.

[12] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid. ALOHA unleashed: A simple recipe for robot dexterity. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=gvdXE7ikHI.

[13] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi:10.15607/RSS.2024.XX.120.

[14] S. Dass, W. Ai, Y. Jiang, S. Singh, J. Hu, R. Zhang, P. Stone, B. Abbatematteo, and R. Martín-Martín. Telemoma: A modular and versatile teleoperation system for mobile manipulation. *arXiv preprint arXiv:2403.07869*, 2024.

[15] Y. Jiang, R. Zhang, J. Wong, C. Wang, Y. Ze, H. Yin, C. Gokmen, S. Song, J. Wu, and L. Fei-Fei. BEHAVIOR robot suite: Streamlining real-world whole-body manipulation for everyday household activities. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=v2KevjWScT.

[16] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi:10.15607/RSS.2024.XX.045.

[17] H. Choi, Y. Hou, C. Pan, S. Hong, A. Patel, X. Xu, M. R. Cutkosky, and S. Song. In-the-wild compliant manipulation with umi-ft, 2026. URL https://arxiv.org/abs/2601.09988.

[18] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. Mahi Shafiullah. Robot utility models: General policies for zero-shot deployment in new environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8275–8283, 2025. doi:10.1109/ICRA55743.2025.11127857.

[19] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, et al. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.

[20] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li, D. Minor, Q. Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

[21] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang. AMO: Adaptive Motion Optimization for Hyper-Dexterous Humanoid Whole-Body Control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi:10.15607/RSS.2025.XXI.061.

[22] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang. CLONE: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=Bw9NHYjDqR.

[23] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025.

[24] C. Mattson, V. Raveendra, E. Novoseller, N. Waytowich, V. J. Lawhern, and D. S. Brown. R2bc: Multi-agent imitation learning from single-agent demonstrations. *arXiv preprint arXiv:2510.18085*, 2025.

[25] K. Song, S. Ma, G. Chen, N. Jin, G. Zhao, M. Ding, Z. Xiong, and J. Pan. Collabot: Vision-language guided simultaneous collaborative manipulation. *arXiv preprint arXiv:2508.03526*, 2025.

[26] W. Zhang, C. Street, and M. Mansouri. Multi-nonholonomic robot object transportation with obstacle crossing using a deformable sheet. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7349–7355, 2025. doi:10.1109/ICRA55743.2025.11128313.

[27] N. Michael, J. Fink, and V. Kumar. Cooperative manipulation and transportation with aerial robots. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009. doi:10.15607/RSS.2009.V.001.

[28] C. Yang, G. N. Sue, Z. Li, L. Yang, H. Shen, Y. Chi, A. Rai, J. Zeng, and K. Sreenath. Collaborative navigation and manipulation of a cable-towed load by multiple quadrupedal robots. *IEEE Robotics and Automation Letters*, 7(4):10041–10048, 2022.

[29] R. T. Fawcett, L. Amanzadeh, J. Kim, A. D. Ames, and K. A. Hamed. Distributed data-driven predictive control for multi-agent collaborative legged locomotion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9924–9930, 2023. doi:10.1109/ICRA48891.2023.10160914.

[30] Y. Wang, M. Damani, P. Wang, Y. Cao, and G. Sartoretti. Distributed reinforcement learning for robot teams: A review. *Current Robotics Reports*, 3(4):239–257, 2022.

[31] B. Wu and C. S. Suh. State-of-the-art in robot learning for multi-robot collaboration: A comprehensive survey. *arXiv preprint arXiv:2408.11822*, 2024.

[32] B. Pandit, A. K. Shrestha, and A. Fern. Multi-quadruped cooperative object transport: Learning decentralized pinch-lift-move. *arXiv preprint arXiv:2509.14342*, 2025.

[33] B. Pandit, A. Gupta, M. S. Gadde, A. Johnson, A. K. Shrestha, H. Duan, J. Dao, and A. Fern. Learning decentralized multi-biped control for payload transport. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=vhGkyWgctu.

[34] J. Zeng, A. M. Gimenez, E. Vinitsky, J. Alonso-Mora, and S. Sun. Decentralized aerial manipulation of a cable-suspended load using multi-agent reinforcement learning. In *2nd Workshop on Safe and Robust Robot Learning for Operation in the Real World*, 2025. URL https://openreview.net/forum?id=yYYmqMv7Al.

[35] K. Shibata, R. Sota, S. D. Bosch, Y. Kadokawa, T. Yoshihisa, and T. Matsubara. Dereco: Decoupling representation and coordination learning for object-adaptive decentralized multi-robot cooperative transport. *arXiv preprint arXiv:2603.08111*, 2026.

[36] S. Chen, Z.-a. Cao, Z. Luo, F. Castañeda, C. Li, T. Wang, Y. Yuan, L. Fan, C. K. Liu, Y. Zhu, et al. Chip: Adaptive compliance for humanoid control through hindsight perturbation. *arXiv preprint arXiv:2512.14689*, 2025.

[37] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, brian ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=vlhoswksBO.

[38] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In J. Tan, M. Toussaint, and K. Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

[39] L. Heng, Y. Tang, J. Xu, H. Bao, D. Huang, and Y. Wang. Humdex: Humanoid dexterous manipulation made easy. *arXiv preprint arXiv:2603.12260*, 2026.

[40] J. Fan, Z. Zhao, Y. Zhang, C. Chen, P. Wang, H. Zhang, and Z. Cheng. Robopaint: From human demonstration to any robot and any view. *arXiv preprint arXiv:2602.05325*, 2026.

[41] M. Shi, S. Peng, J. Chen, H. Jiang, Y. Li, D. Huang, P. Luo, H. Li, and L. Chen. Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. *arXiv preprint arXiv:2602.10106*, 2026.

[42] R. Nai, B. Zheng, J. Zhao, H. Zhu, S. Dai, Z. Chen, Y. Hu, Y. Hu, T. Zhang, C. Wen, et al. Humanoid manipulation interface: Humanoid whole-body manipulation from robot-free demonstrations. *arXiv preprint arXiv:2602.06643*, 2026.

[43] H. Chen, W. Zhang, P. Li, S. Ma, K. Ma, Y. Jin, Z. Xu, X. Wang, Y. Zheng, Z. Wang, et al. Rhythm: Learning interactive whole-body control for dual humanoids. *arXiv preprint arXiv:2603.02856*, 2026.

[44] W.-J. Huang, Y.-Y. Zhang, Y.-L. Wei, Z.-W. Xia, J. Tan, Y.-M. Li, Z. Zhao, and W.-S. Zheng. Learning whole-body human-humanoid interaction from human-human demonstrations. *arXiv preprint arXiv:2601.09518*, 2026.

[45] J. Mao, S. Zhao, S. Song, C. Hong, T. Shi, J. Ye, M. Zhang, H. Geng, J. Malik, V. Guizilini, and Y. Wang. Universal humanoid robot pose learning from internet human videos. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 1–8, 2025. doi:10.1109/Humanoids65713.2025.11203143.

[46] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

[47] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[48] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[50] R. Cadene, S. Alibert, A. Soare, Q. Gallouedec, A. Zouitine, S. Palma, P. Kooijmans, M. Aractingi, M. Shukor, D. Aubakirova, M. Russi, F. Capuano, C. Pascal, J. Choghari, J. Moss, and T. Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

Contents

A Dual Robot Teleoperation System

A.1 Extended Teleoperation Implementation Details

Building upon the dual-robot teleoperation interface introduced in Section 3.1, this section provides additional technical details regarding system-wide control and transmission frequencies, lower-level smoothing algorithms, and hand-retargeting mechanics. The teleoperation interface relies on a pair of PICO4 Ultra VRs, which capture operators' motions as full-body human pose estimates in SMPL format using the headsets, handheld controllers, and two pairs of ankle trackers. The raw full-body human poses are retargeted and continuously transmitted to both robots at 50 Hz. Once processed, these retargeted joint-space commands are executed with a control frequency of 500 Hz.

Because raw human movement contains high-frequency jitter that can destabilize physical hardware, each robot's lower-level controllers process this stream differently. For G1, the SONIC [20] framework manages these high-frequency commands to maintain whole-body balance and stability. Conversely, for the Vega1 mobile manipulator, we engineered a custom lower-level inference controller that applies Exponential Moving Average (EMA) smoothing to the command stream using the update rule $S_t = \alpha x_t + (1 - \alpha)S_{t-1}$, where S_t and S_{t-1} represent the current and previous smoothed actions, x_t is the raw incoming command, and the smoothing factor α is set to 0.01.

While G1's hand actuation is natively managed within SONIC, we designed a simplified hand retargeting approach for the Vega1 mobile manipulator. Similar to TWIST2 [23], instead of continuous hand pose estimation, we treat the F5D6 hand of Vega1 entirely as a gripper. By decoupling complex finger tracking from the interface, operators control the hand's actuation cleanly and reliably via button presses on the PICO handheld controllers. The VR controllers capture the operator's gripping intent as a continuous signal, which the host software translates into a normalized scalar grasp command, $\gamma \in [0, 1]$. In this mapping, $\gamma = 0$ denotes a fully open hand and $\gamma = 1$ represents a closed hand. The controller relies on two predefined joint configurations: an open pose (q_{open}) and a close pose (q_{close}). The final commanded hand joint configuration is continuously computed via linear interpolation: $q_{\text{hand}} = (1 - \gamma)q_{\text{open}} + \gamma q_{\text{close}}$.

A.2 Dual-Operator Interface Mapping

To empower two operators to fully manage the heterogeneous dual-robot teleoperation system independently, eliminating the need for a dedicated host PC supervisor, we designed a mapping of the controller inputs to all essential robotic and system-level actions (detailed in Table 2).

A.3 Rationale for Retargetting Frameworks

The General Motion Retargeting (GMR) [46] algorithm provides a highly effective, universal mapping solution for cross-embodiment control. Because the Vega1 mobile manipulator requires only upper-body articulation on a stable wheeled base, we purpose-built its control pipeline to leverage GMR. However, direct kinematic retargeting is insufficient for the G1 humanoid. Retargeting human motion to the G1 humanoid requires an extra layer of control to ensure lower-body stability and high-fidelity physical resemblance, which we achieve by utilizing SONIC [20], the state-of-the-art framework for physics-based, whole-body teleoperation for the G1 humanoid.

A.4 Robot Data Collection

To facilitate temporal alignment across domains, clip durations are fixed on a per-task, per-domain basis. Specifically, recording durations are fixed on a strictly per-task basis. For any given task, all robot teleoperation data clips share a uniform duration, while all human-human demonstrations

Task	G1 Controller Mapping	Vega1 Controller Mapping
Toggle Teleoperation (Start/Pause)	Menu	X
Record Data	Right Grip + A	Y
Reset to Initial Pose and Pause	A + X	A
Camera View (Zoom in/Out)	B	B
Trigger	Close Hand	Close Hand
Grip	Open Hand	Open Hand
Keep Recorded Data	Left Grip + X	Left Axis Click
Yaw Control (ω_z)	N/A	Left Axis
Velocity Control (v_x, v_y)	N/A	Right Axis

Table 2: **Controller Mapping.** Comprehensive input mapping detailing how operator commands are translated into specific robotic behaviors and system operations for both the G1 and Vega1 robots.

share a standardized duration that is proportionally reduced by a factor of 1.5, as introduced in Section 4.1. Clip length for each task are detailed in Table 4. This approach is motivated by the need to enforce synchronized motion representations [7], which empirically improves velocity predictions and preserves motion smoothness when incorporating human pretraining priors. To align the dual robot data streams offline, we rely on common timestamps synchronized via a centralized host computer running a Chrony NTP server. This post-alignment procedure allows us to accurately synchronize the data from both robots, resulting in a cohesive 10 Hz dataset comprising time-aligned egocentric images, joint-space poses, and Vega1 base velocities.

A.5 Teleoperation System Example

As shown in Figure 8, this example demonstrates our teleoperation system's ability to successfully capture complex human tasks and translate them directly to the robots in real time. The figure illustrates the synchronized execution of a collaborative loco-manipulation task alongside the operators.

To provide further details regarding the amortized collection times shown in Figure 5, we outline the specific operational protocols used for each setup. To optimize the robotic data gathering process, tasks were executed by our most experienced teleoperator, who was occasionally supported by a third-party assistant dedicated to rapidly resetting the workspace between trials. The human-human demonstration data was collected by an arbitrary demonstrator operating entirely independently, managing all tasks and environment resets without external assistance.

B Human Data Collection Pipeline

In this section, we detail the human data collection with hyperparameter choices listed in Table 5.

B.1 Video capture and alignment

Multi-view RGB-Depth capture uses an array of Intel RealSense D-series cameras, where one camera is the master and the rest two placed on head/neck mounts of the humans. The master camera captures the two humans in its frame during the task. Each camera streams color and depth at $W \times H = 640 \times 480$ at 15 Hz. Each clip is consistently T seconds long for each task. During recording, the depth frame passes through a spatial, a temporal, and a hole filling filter. For each camera c_i (where c_m denotes the master camera and c_1, c_2 denote the egoview cameras), we also capture the camera intrinsics including focal lengths $(f_x^{c_i}, f_y^{c_i})$ and the principal points $(c_x^{c_i}, c_y^{c_i})$. After recording, we perform a nearest-neighbor alignment between the cameras.

[Real-Time Teleoperation] Vega1 and G1 stand on opposite sides of a table **(Initial)**. Both robots approach the table **(#1)**. G1 reaches out and grabs the toy dog **(#2)**. G1 then lifts the toy, and Vega1 extends its arm to receive the handover **(#3)**.

[Real-Time Teleoperation] As Vega1 takes hold of the toy, G1 begins to let go **(#4)**. Once Vega1 secures its grip, G1 fully releases the object **(#5)**. Vega1 then turns around, approaches the long grey table **(#6)**, and places the toy on the table **(#7).**

Figure 8: **Teleoperation System Example.** A side-by-side of the human teleoperator and the synchronized dual-robot execution. The sequence demonstrates the G1 humanoid walking to the workspace to grasp the target object (a toy dog), followed by a coordinated physical handover to the Vega1 mobile manipulator, which subsequently pivots and places the object onto a secondary table.

B.2 Mesh Extraction

From each clip's aligned master-camera RGB streams, we offline extract per-frame, per-person 3D anatomical skeletons and body meshes through a two-stage detect-then-fit pipeline. In **the detection and tracking stage**, every color frame is passed through a YOLOv8-nano [47] detector restricted to the COCO person class. The detections are streamed through the ByteTrack multi-object tracker to track identifiers across frames. A semantic identity assignment is performed once and frozen for the remainder of the clip: the active tracks are sorted in ascending order of bounding-box center x coordinate, allowing training-time correspondence binding each role (G1, Vega1) to a fixed demonstrator. For invalid detections, we reuse bounding boxes from previous frames optionally. In the **mesh fitting stage**, the bounding boxes are fed into the SAM 3D Body DINOv3 [10] estimator, which returns per person and per frame, (i) the 3D anatomical keypoints $\mathbf{K}_t^{(p)} \in \mathbb{R}^{70\times 3}$ in the mhr70 convention (ii) the dense body-mesh vertices $\mathbf{V}_t^{(p)} \in \mathbb{R}^{N_V\times 3}$; (iii) the per-keypoint 2D image-plane projection $\mathbf{u}_t^{(p)} \in \mathbb{R}^{70\times 2}$ used by the calibration stage, (iv) the predicted camera translation $\mathbf{c}_t^{(p)} \in \mathbb{R}^3$ that places the SMPL-style canonical body inside a perspective camera; and (v) the predicted focal length $f_t^{(p)} \in \mathbb{R}^+$ of that perspective model. The raw-pose tensors live in the pseudo-camera frame of the SAM 3D Body model rather than the master camera frame, and thus we carry out a depth-based calibration step.

B.3 Depth-based Calibration

For calibration, every frame t and every semantic identity p is processed as follows. We first lift the SAM keypoints and mesh vertices out of the model's pose only canonical frame into its perspective

camera-frame by adding the predicted camera translation

$$\widetilde{\mathbf{K}}_t^{(p)} = \mathbf{K}_t^{(p)} + \mathbf{c}_t^{(p)}, \widetilde{\mathbf{V}}_t^{(p)} = \mathbf{V}_t^{(p)} + \mathbf{c}_t^{(p)}.$$

The lifted mesh vertices are then projected through the SAM perspective intrinsics $(f_t^{(p)}, f_t^{(p)}, W/2, H/2)$ onto the image plane and aggregated into a per-pixel z-buffer retaining the minimum z at each pixel as the implicit depth of the body surface viewed from the SAM perspective camera viewpoint: Concretely, for each vertex $\mathbf{v}_k = (x_k, y_k, z_k) \in \widetilde{\mathbf{V}}_t^{(p)}$ the per-pixel projection is

$$u_k = \text{round}\left(f_t^{(p)} \cdot \tfrac{x_k}{z_k+\varepsilon} + \tfrac{W}{2}\right), \qquad v_k = \text{round}\left(f_t^{(p)} \cdot \tfrac{y_k}{z_k+\varepsilon} + \tfrac{H}{2}\right),$$

with $\varepsilon = 10^{-9}$, and the z-buffer aggregates only the closest vertex per pixel,

$$z_{\text{sam}}(u,v) = \min\{z_k : (u_k, v_k) = (u,v), z_k > 10^{-6}, 0 \le u_k < W, 0 \le v_k < H, \},$$

with $z_{\text{sam}}(u,v) = +\infty$ at pixels onto which no in-bounds vertex projects. The observed metric depth, $z_{\text{obs}}(u,v)$, from the master RealSense sensor is recovered from the captured depth image. The two depths are compared inside an eroded person mask $\mathcal{M}_t^{(p)}$, the per-frame bounding-box rectangle shrunk by a 5×5 kernel for one iteration to discard boundary pixels likely to straddle the body silhouette and the background. Letting $\mathcal{P}_t^{(p)}$ denote the pixels satisfying the joint validity criterion

$$(u,v) \in \mathcal{M}_t^{(p)} \wedge 10^{-6} < z_{\text{sam}}(u,v) < +\infty \wedge \text{depth image}(u,v) > 0,$$

an initial scale factor is recovered as the element-wise median of observed-to-rendered depth ratios,

$$s_0^{(p)}[t] = \text{median}_{(u,v)\in\mathcal{P}_t^{(p)}} \frac{z_{\text{obs}}(u,v)}{\max\left(z_{\text{sam}}(u,v), 10^{-8}\right)}.$$

The set $\mathcal{P}_t^{(p)}$ is required to contain at least 500 pixels; if the eroded mask is too aggressive, the criterion is relaxed by replacing $\mathcal{M}_t^{(p)}$ with the un-eroded rectangle. Given a valid initial estimate, we sharpen it with a least-squares refit on the inlier subset defined by twin tolerance bands around the predicted depth $\hat{z}_{\text{sam}}(u,v) = s_0^{(p)}[t] \cdot z_{\text{sam}}(u,v)$:

$$\mathcal{I}_t^{(p)} = \{(u,v) \in \mathcal{P}_t^{(p)} : z_{\text{obs}}(u,v) \ge \hat{z}_{\text{sam}}(u,v) - \tau_{\text{front}} \wedge |z_{\text{obs}}(u,v) - \hat{z}_{\text{sam}}(u,v)| < \tau_{\text{abs}}\},$$

with $\tau_{\text{front}} = 0.10$m, allowing the observed depth to lie up to 10 cm in front of the rendered surface and thereby absorbing thin self-occluders, and $\tau_{\text{abs}} = 0.30$m, rejecting any pixel whose depth disagrees by more than 30 cm. Provided $|\mathcal{I}_t^{(p)}| \ge 20$, the refined scale is the closed-form linear-least-squares ratio

$$s^{(p)}[t] = \frac{\sum_{(u,v)\in\mathcal{I}_t^{(p)}} z_{\text{obs}}(u,v) \cdot z_{\text{sam}}(u,v)}{\max\left(\sum_{(u,v)\in\mathcal{I}_t^{(p)}} z_{\text{sam}}(u,v)^2, 10^{-12}\right)},$$

which replaces $s_0^{(p)}[t]$; otherwise the initial median estimate is kept unchanged. Letting $\widetilde{z}_{t,j}^{(p)}$ denote the z-component of the lifted keypoint $\widetilde{\mathbf{K}}_{t,j}^{(p)}$, the recovered scale converts it into RealSense-metric units,

$$z_{t,j}^{(p)} = s^{(p)}[t] \cdot \widetilde{z}_{t,j}^{(p)}.$$

We then anchor each keypoint into the master color camera's intrinsic frame by treating the SAM-predicted 2D image-plane projection $\mathbf{u}_t^{(p)} = (u_{t,j}, v_{t,j})_{j=1}^{70}$ as observed pixel coordinates in the RealSense color image and $z_{t,j}^{(p)}$ as the corresponding metric depth, and back-projecting through the master color pinhole as

$$\mathbf{p}_{t,j}^{(p)} = \left(\frac{(u_{t,j} - c_x^{c_m}) z_{t,j}^{(p)}}{f_x^{c_m}}, ; \frac{(v_{t,j} - c_y^{c_m}) z_{t,j}^{(p)}}{f_y^{c_m}}, z_{t,j}^{(p)}\right).$$

A final axis re-labeling rotates the result from the RealSense optical convention $(X_{\text{right}}, Y_{\text{down}}, Z_{\text{forward}})$ into the visualization/world convention $(X_{\text{forward}}, Y_{\text{left}}, Z_{\text{up}})$ used everywhere downstream. We use the same notation $\mathbf{p}_{t,j}^{(p)}$ to denote the final depth calibrated poses without loss of generality. The calibrated keypoint trajectories contain jitter from depth noises. We address this with a moving average applied independently per person, per joint, and per coordinate, awaring the validity of the extracted keypoints.

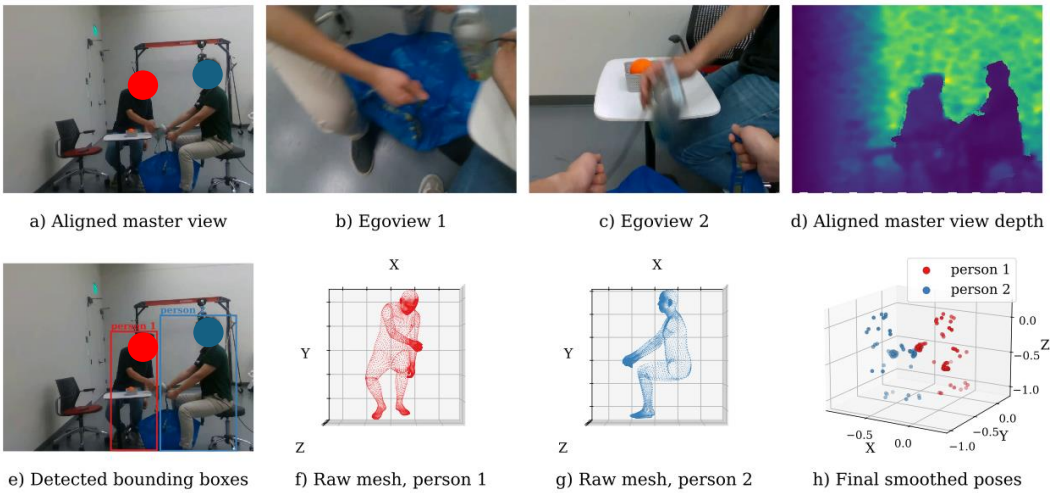


Figure 9: **Human Data Collection Example.**

B.4 Human Data Collection Example

As shown in Figure 9, this example shows the capacity of our human data collection pipeline in retargeting human data for collective training. We in sequence show example aligned master and egoview camera images (together with the master view depth) as outcome of Section B.1, example bounding boxes and meshes from Section B.2 and example extracted poses $\mathbf{p}_t^{(p)}$ after smoothing from Section B.3. The poses in Figure 9 (h) are recentered so that the position of person 1's neck at the first valid frame of the clip is the origin.

C Data Preprocessing and Model Architecture

In this section, we introduce our model architecture, along with how we process the raw data. See our choice of hyperparameters in Table 6.

C.1 Data Preprocessing

We introduce our data preprocessing. During training, our model takes in a K-step history of observations. On the robot side, we resize the historic image stream into size $(K, 2, 224, 224, 3)$ where 2 robots each input K steps of egoview RGB images of size $224 \times 224 \times 3$. In addition to the image stream, the policy is conditioned on a K-step history of poses, generated via forward kinematics using each robot's URDF and recorded joint states. We recenter the poses in root-relative coordinates by subtracting the position of each robot's head joint. From the full skeleton, we retain a fixed subset of 9 joints per robot. For G1, we use the keypoints in sequence of head, left wrist pitch, right wrist pitch, left elbow, right elbow, left shoulder pitch, right shoulder pitch, left hip pitch, and right hip pitch. For Vega1, we use the keypoints in sequence of head joint 3, left arm joint 7, right arm joint 7, left arm joint 4, right arm joint 4, left arm joint 1, right arm joint 1, and we repeat the entries of torso joint 3 twice to match the input dimensions with the human input. For Vega1, we additionally include a velocity pseudo-joint that encodes the mobile base's planar velocity command (v_x, v_y, ω_z) as a 3-vector. Together, the K-step history of poses is of shape $(K, 57)$ where each keypoint pose is a 3-dimensional vector. We flatten this high-dimensional pose vector to $K \times 57$ before being projected into the transformer's hidden dimension. The ground truth output of the model is a chunk of length H consisting of a future pose and a future joint-space (qpose) block. The pose block mirrors the structure of the state stream exactly. The qpose block consists of 23 joint angles from G1 and 18 from Vega1. Unlike the pose block, qpose is never observed as part of the policy's input; it appears only as a prediction target on the output side, providing low-level command signal that ultimately drives the physical actuators. Together, the shape of the prediction target is $H \times 98$.

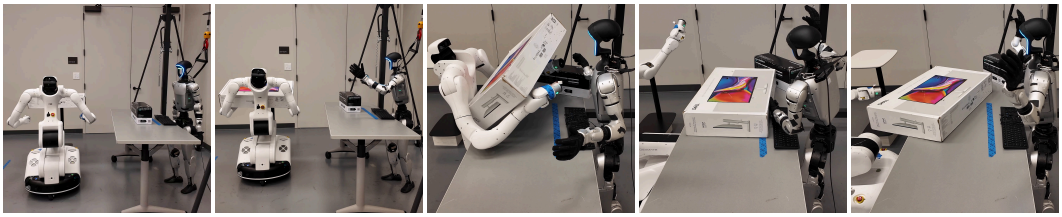
On the human side, demonstrations are provided as per-clip 3D keypoint trajectories on the mhr70 skeleton expressed in a shared world frame (See Section B). Each human represents a robot. The two egoview RGB streams associated with assigned camera serials are also resized to 224×224 to match the robot side. For each of the two designated persons, every keypoint position has the neck position subtracted, in direct analogy to the robot side. The camera's frame is not naturally aligned with the kinematics frame that the robots induce, so for each human we apply a single fixed rotation matrix $R_{\text{role}} \in SO(3)$, parameterized by intrinsic XYZ Euler angles in degrees and applied to every keypoint at every timestep. Each R_{role} is chosen by inspecting the subject's orientation in the first frame of a representative clip and is held constant across all frames. It is not re-estimated per frame and therefore does not track the subject's rotation across each clip. This optional design choice, however, empirically shows that human priors are largely rotation-invariant. From the 70 mhr70 keypoints, we retain the same nine anatomical joints for both persons: neck, left and right wrists, left and right elbows, left and right shoulders, and left and right hips. For the person representing Vega1, we additionally compute a velocity pseudo-joint (v_x, v_y, w_z) from finite differences of the rotated but uncentered pelvis position (taken as the midpoint of the left and right hip keypoints), where ω_z is the finite difference of the inter-hip vector's azimuthal angle. The resulting vector is multiplied by a scalar of 0.667 so that the magnitude matches the dexmate odometry to account for the differences in human vs. robot speed during data collection. In summary, the K-step pose history is of shape $(K, 57)$ (flattened for projection into the transformer's hidden dimension). The accompanying K-step image history is of shape $(K, 2, 224, 224, 3)$. Unlike the robot side, the ground truth prediction target is a chunk of length H consisting of only the pose block of size $H \times 57$. The qpose block is absent because human demonstrations do not carry joint-angle measurements.

All inputs and targets are standardized before being projected into the transformer. Following [7], we compute separate normalization statistics for the robot and human data. For the image stream, each frame is first rescaled from the discrete pixel range to $[0, 1]$, then standardized per channel using the ImageNet [49] statistics $\boldsymbol{\mu}\text{img} = (0.485, 0.456, 0.406)$ and $\boldsymbol{\sigma}\text{img} = (0.229, 0.224, 0.225)$, applied as $\mathbf{x} \mapsto (\mathbf{x} - \boldsymbol{\mu}\text{img})/\boldsymbol{\sigma}\text{img}$ over the channel axis. For the state stream and the action target, we apply per-dimension z-score standardization with mean and standard deviation estimated empirically on the training split of each modality.

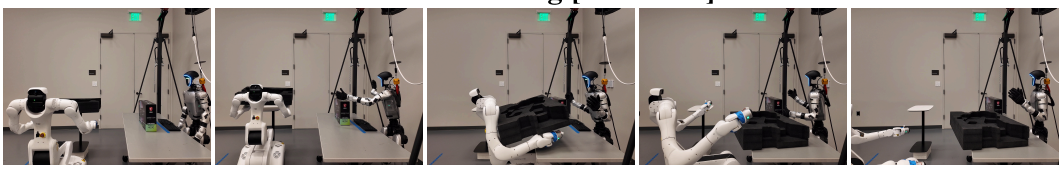
C.2 Architecture Detail

The main architecture is implemented with [50] following a standard ACT procedure [11] (Figure 4). We start with a description of the training only conditional variational autoencoder (CVAE) branch, which encodes the ground truth action chunk together with the proprioceptive state into a latent code that conditions the main decoder. The reconstruction target is truncated to the pose block so that the latent captures only structure shared across the robot and human modalities. The encoder consumes three input streams, each lifted into the hidden dimension $D = 512$ by a dedicated projector, assuming a batch size of B: a learnable CLS aggregator broadcast to $(B, 1, D)$; the K-step pose projected by a $\mathbb{R}^{K \cdot 57} \to \mathbb{R}^D$ linear projector, and the pose-only action chunk projected per timestep by a linear projector to H tokens of shape (B, H, D). These tokens are concatenated into a length $H + 2$ sequence, summed with a fixed sinusoidal positional embedding, masked along the action axis, and processed by a four-layer post-norm transformer, producing a CLS summary that is linearly projected and split into the mean and log-variance of a 32-dimensional Gaussian regularized against the unit-Gaussian prior.

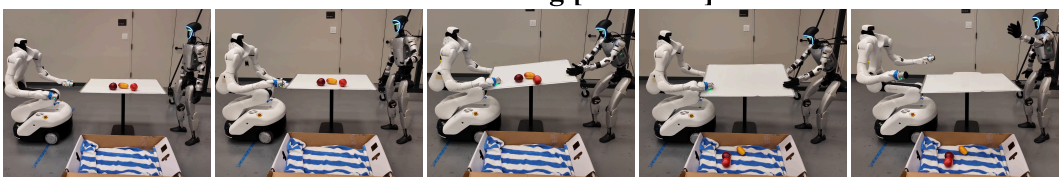
The main transformer encoder consumes three token streams: the CVAE latent z (sampled at training and zeroed at inference); the K-step poses, projected to one token $(B, 1, 512)$ via a linear projector; and the egoview RGB images, each passed through an ImageNet-pretrained ResNet-18 backbone, mixed by a 1×1 convolution. A learnable embedding over the latent and pose tokens and a fixed 2 dimensional sinusoidal position embedding over the image tokens are added at every layer before passing through a four-layer post-norm transformer encoder with 8 heads and FFN width 3200. The single-layer main transformer decoder takes a zero content input of shape $(H, B, 512)$ and a learned positional embedding for DETR-style object queries. The decoder layer applies self-attention, cross-attention onto the encoder memory, and a $512 \to 3200 \to 512$ FFN, followed by a final layer norm.



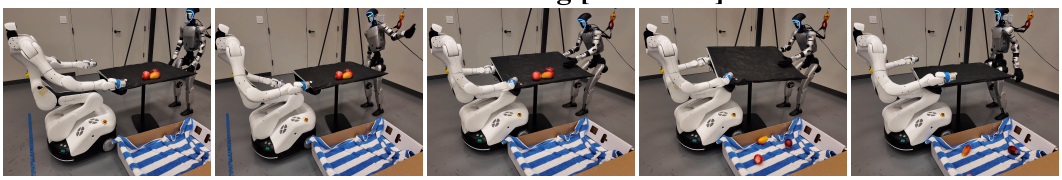
T2: Box Moving [Standard]



T2: Box Moving [Modified]



T3: Board Tilting [Standard]



T3: Board Tilting [Modified]

Figure 10: **Zero-Shot Generalizability to OOD Conditions.** We evaluate DUET under visual and physical distribution shifts. In **T2**, the standard boxes are replaced with a black foam box and a green-and-black box, altering visual cues and physical dynamics. In **T3**, the white board is entirely covered by a black cover to test resilience to drastic background variations. These changes are illustrated in the figure. Despite these perturbations, the DUET framework successfully coordinates the execution to complete the tasks.

For robot batches, the decoder output is passed through a sibling pair of three-layer MLPs, each $512 \to 512 \to 512 \to d_{\text{out}}$ with intermediate ReLUs, producing pose ($d_{\text{out}} = 57$) and qpose predictions ($d_{\text{out}} = 41$) respectively. For human samples, the qpose branch does not exist, receiving no gradient during human-only pretrain.

In Stage 1, we pretrain the entire architecture with all weights unfrozen using the human data. In Stage 2, we warm-start all weights except the qpose MLP (cold-started) and finetune on robot data. At inference, all blocks run identically with the trained weights with CVAE style latent set to zero.

D Real-World Experiment Details

D.1 Detail Evaluation Metrics

As introduced in Section 4.3, we implement a normalized metric scoring for each task. For **T1: Trash Collection**, 0.5 points is awarded if any single piece of trash is successfully deposited into the target basket **[Single Item]**, and the full 1.0 point is achieved when both pieces are secured **[Both Items]**. For **T2: Box Organization**, 0.5 points are awarded when G1 clears the workspace while Vega1 successfully secures the box **[Clear & Secure]**, and the remaining 0.5 points are granted if the box is placed on the table without crossing the boundary line, which includes successful recoveries via a corrective nudge from G1 **[In-Bound Placement]**. For **T3: Board Tilting**, 0.5

points are awarded if the white board remains on-table and at least one of the three items lands in the basket **[Partial Containment]**, with the full 1.0 point achieved only when all three items are safely contained **[Full Containment]**. For **T4: Doll Passing**, the initial object grasp by Vega1 **[Initial Grasp]** and the subsequent dual-robot handover **[Successful Handover]** each contribute 0.5 points. The detailed score distributions and aggregate success rates are summarized in Table 3.

T1 : Trash Collection	Single Item	Both Items	Points ↑	Success Rate ↑
DUET (Ours)	9	4	6.5	4/10
Robot-only (50)	9	4	6.5	4/10
Robot-only (30)	10	3	6.5	3/10
T2: Box Organization	Clear & Secure	In-Bound Placement	Points ↑	Success Rate ↑
DUET (Ours)	10	6	8	6/10
Robot-only (50)	9	5	7	5/10
Robot-only (30)	8	3	5.5	3/10
T3 : Board Tilting	Partial Containment	Full Containment	Points ↑	Success Rate ↑
DUET (Ours)	10	7	8.5	7/10
Robot-only (50)	7	6	6.5	6/10
Robot-only (30)	6	3	4.5	3/10
T4: Doll Passing	Initial Grasp	Successful Handover	Points ↑	Success Rate ↑
DUET (Ours)	8	3	5.5	3/10
Robot-only (50)	9	2	5.5	2/10
Robot-only (30)	7	1	4	1/10

Table 3: **Real-World Benchmarking Results.** Detailed evaluation reporting partial sub-task progress and overall success rates and points. A trial is considered successful only when all constituent sub-tasks are completed. Blue highlights denote the DUET performance in each category.

D.2 Deployment

During real-world deployment, the centralized inference module monitors the data streams from both the G1 and Vega1. The module takes in real-time egocentric camera images alongside qposes from both robots, which are mapped online via Forward Kinematics to generate synchronized poses. These poses are then transformed into relative head frames with respect to each robot to ensure consistency, according to Section C.1. Based on these inputs, the policy predicts target qpose for both robots and base velocity commands for the Vega1 at a fixed frequency of 10 Hz. To guarantee operational safety and preserve multi-robot coordination, we enforce a synchronization protocol: the 10 Hz action commands are only sent if inputs are simultaneously received from both robots. In the event of a asynchronous delay, the system defaults to a zero-order hold, freezing both robots at their last executed actions to prevent uncoordinated movement. Finally, the commanded actions are executed by the hardware via the same low-level control loops utilized during teleoperation (Section A.1), which run independently at 500 Hz to ensure stable, robust, high-frequency action.

D.3 Generalizability

To evaluate the robustness and zero-shot generalizability of the DUET framework, we designed a series of out-of-distribution (OOD) experiments focusing specifically on **T2: Box Organization** and **T3 : Board Tilting**. For T2, we introduced perturbations by replacing the standard white box with a black foam box and altering the G1 robot's manipulation white-and-black box to a green-and-black box. In T3, we obscured the white board with a black cover to drastically shift the background visual distribution. As illustrated in Figure 10, our framework successfully adapts to these severe variations. Crucially, these object substitutions alter not only the visual cues but also the underlying physical dynamics, such as material friction during contact.

E Hyperparameters

This section details the hyperparameters used throughout our data collection and training phases. Task-specific clip lengths for robotic and human data are summarized in Tables 4 and 5, respectively. Additionally, Table 6 outlines the core training hyperparameters. All models were trained using a single NVIDIA A100-SXM4-80GB GPU per checkpoint.

Parameter	**Value**
T1: Trash Collection Clip Length (T)	30 seconds
T2: Box Organization Clip Length (T)	60 seconds
T3: Board Tilting Clip Length (T)	20 seconds
T4: Doll Passing Clip Length (T)	30 seconds

Table 4: Robot Data Recording Parameters

Parameter	**Value**
T1: Trash Collection Clip Length (T)	20 seconds
T2: Box Organization Clip Length (T)	40 seconds
T3: Board Tilting Clip Length (T)	13 seconds
T4: Doll Passing Clip Length (T)	20 seconds

Table 5: Human Data Recording Parameters

Task	$R_{\text{role}}^{\text{G1}}$ (Degrees)	$R_{\text{role}}^{\text{Vega1}}$ (Degrees)	B	K	H
T1	$(0, 0, 180)$	$(0, 0, -90)$	32	5	400
T2	$(0, 0, 90)$	$(0, 0, 180)$	32	3	40
T3	$(0, 0, -90)$	$(0, 0, 90)$	32	3	40
T4	$(0, 0, 90)$	$(0, 0, 180)$	32	5	400

Table 6: **Training Parameters.** $R_{\text{role}}^{\text{G1}}$ and $R_{\text{role}}^{\text{Vega1}}$ refer to the rotation matrices described in Section C.1. B refers to the training batch size, K refers to the history length, and H refers to the future predicted chunk length. The same parameters are used consistently across the 30 data robot-only, the 50 data robot-only, the 60 data human pretrained and the 30 robot data finetuned checkpoints.