# 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

***Abstract*— Data acquisition is a bottleneck in training dual-robot collaboration systems. We introduce DUET, an efficient dual-robot learning framework that uses human coordination priors for mobile manipulation. We introduce a VR-based dual-robot teleoperation system alongside a two-human pose tracking pipeline. By pretraining an Action Chunking Transformer on human demonstrations and finetuning on minimal real-robot data, DUET eliminates the need for massive teleoperation datasets. Evaluated on four tasks using a Unitree G1 and a Dexmate Vega1, DUET achieves equal or superior performance compared to baselines trained with robot-only data. Our project page is available at https://zhaoy37.github.io/Duet/.**

## I. INTRODUCTION

Many real-world tasks intrinsically exceed the kinematic reach and payload limits of a single robot. Yet, realizing the practical utility of dual-robot systems requires advancing toward contact-rich collaborative mobile manipulation [2]. Since modeling the multi-body dynamics of such joint operations analytically is computationally prohibitive for general purpose, end-to-end visuomotor policies offer a compelling alternative. These policies, however, are severely bottlenecked by data acquisition. Orchestrating teleoperation for heterogeneous dual-robot systems demands complex setups and multiple human experts, making large-scale real-robot data collection costly and slow.

We design a dual-operator teleoperation framework that allows two humans to efficiently command a heterogeneous duo. While this is a step forward in high-fidelity data collection, teleoperation inherently remains limited by time and cost. We complement our system with direct human-human demonstrations [3], which are faster to capture, decouple data volume from hardware constraints, and naturally encode rich coordination priors. To synthesize these modalities, we introduce DUET (**D**ual-robot **U**nderstanding via **E**fficient **T**eaching), a centralized visuomotor policy framework for heterogeneous dual-robot collaboration.

DUET integrates a dedicated two-operator teleoperation pipeline with efficient human demonstrations harnessed via SAM 3D Body [4]. By projecting all data into a shared pose space, a single Action Chunking Transformer (ACT) [5] backbone pretrains on low-cost human data and finetunes on a minimal set of high-fidelity robot data. To evaluate this collective training architecture, we introduce a novel benchmark of four contact-rich collaborative tasks. Validated on a Unitree G1 and a Dexmate Vega1, DUET shows that integrating human priors lowers the teleoperation overhead while achieving equal or superior task performance compared to fully robot-demonstrated baselines.

[†]Equal contribution. Order is determined by coin flip.

[‡]Corresponding author: yiqizhao@usc.edu

[‡]This paper is an extended abstract of our full paper [1].

## II. RELATED WORK

While imitation learning via teleoperation has successfully scaled single-robot mobile manipulation [2], extending high-fidelity data collection for spatially distributed, heterogeneous multi-robot teams remains unresolved. Furthermore, existing multi-robot collaboration heavily relies on classical control or reinforcement learning [6], which struggle to generalize in unstructured, visually rich environments. We are thus motivated to harness end-to-end vision-action models with high generalizability for real-world manipulation tasks. Recent visuomotor policies have leveraged human demonstrations as structural priors to bypass the bottleneck of slow on-robot data collection [3]. However, 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.

## III. METHOD

### A. Teleoperation Pipeline

We introduce a 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 via 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-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 use PICO VR headsets paired with handheld controllers and 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 $1920 \times 1080$ (30 Hz) stream from a RealSense camera, while the Vega1 operator receives a wide-angle

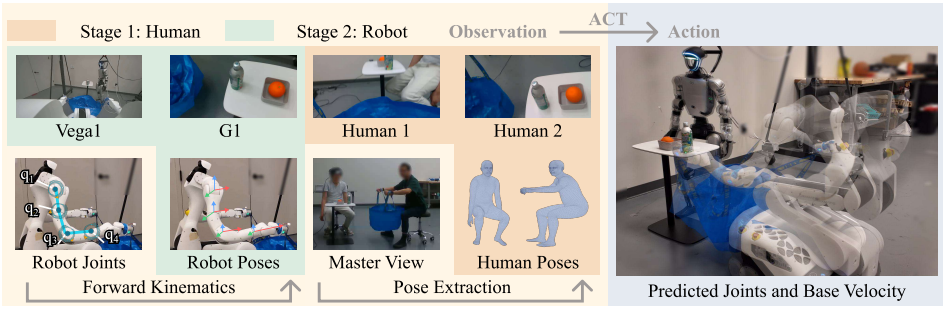



Fig. 1: **DUET Overview.** In *stage* 1, we pretrain an ACT architecture (Figure 2) with RGB streams from human-human demonstrations and extracted keypoints for proprioceptives. In *stage 2*, in-distribution robot data finetunes the ACT policy.

$2560 \times 720$ (30 Hz) stream from a ZED camera. To eliminate the 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, 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 [7] 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) [8] 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 $\omega_z$. 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.

### B. 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 [9] human detector to produce a set of 2 bounding boxes tracking the human identities. For each human, SAM 3D Body [4] 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.

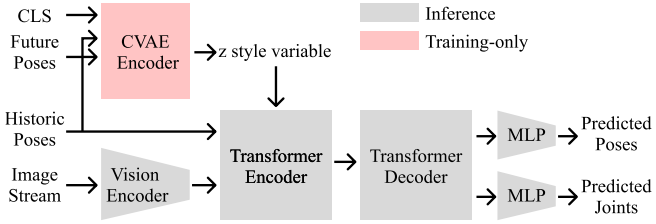



Fig. 2: **Model Architecture.** Our ACT policy uses a 2-head design with the joint head randomly initialized at finetune.

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 $s_0 = \text{median}_{(u,v)\in\mathscr{M}} \frac{z^{\text{obs}}(u,v)}{\hat{z}(u,v)}, s = \frac{\sum_{(u,v)\in\mathscr{I}} z^{\text{obs}}(u,v)\,\hat{z}(u,v)}{\sum_{(u,v)\in\mathscr{I}} \hat{z}(u,v)^2}$, where $\mathscr{M}$ is the eroded silhouette and $\mathscr{I} \subseteq \mathscr{M}$ 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. A per-keypoint mask-weighted moving-average filter suppresses high-frequency jitter introduced by depth-sensor noise. We extract 9 3D keypoints per human and root velocities from pelvis movements, which we then use to pretrain our policy.

### C. Collective Training

Our robots are controllable via *joint-space* commands. Our architecture (Figure 2) is inspired by [3] and we train a single policy in two stages that share an ACT [5] based architecture but differ in the data source. As depicted in Figure 1, **Stage 1** pretrains the policy backbone on human-human collaboration data collected following Section III-B; **Stage 2** finetunes the resulting weights on robot teleoperation data of Section III-A. 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 use 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 $\hat{p}$ for the prediction of ground-

| 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 |

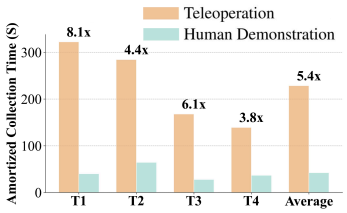



Fig. 3: **Benchmark data distribution and collection efficiency. Left:** Total data span and number of episodes per pipeline across four tasks. **Right:** *Amortized collection time* (AT) for each task. The value above each task is the *speedup ratio*: the teleoperation $AT$ divided by the human-demonstration $AT$, where $AT$ is the average time for each successful data collection.

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 III-B 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 III-B and estimate the base velocity from pelvis.

We utilize an ACT policy, 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 [3] 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 [5], 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.

## IV. 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 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.

### A. Benchmark and Evaluation Setup

We introduce a benchmark suite of four tasks designed to isolate key multi-robot challenges. For each task, we use our teleoperation pipeline (Sec. III-A) to collect high-fidelity robot data on a heterogeneous platform, comprising Vega1 and G1, alongside corresponding human-human demonstrations (Sec. III-B): **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 3. 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 [3]. 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.

### B. 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 3 (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, human demonstration provides a **5.4**$\times$ acceleration on average. This underscores the advantages of

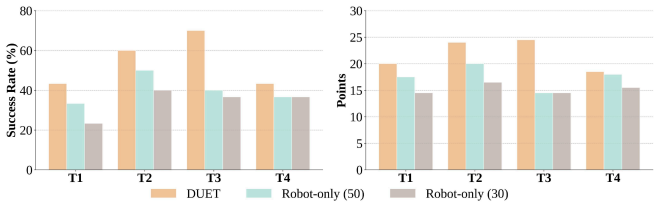



Fig. 4: **Evaluation Results.** Task success rates and accumulated points over 30 experiments for each task. DUET (60 human demos + 30 robot trajectories) is compared against robot-only baselines (trained on 30 and 50 trajectories).

human data collection: it improves collection efficiency while bypassing the high capital expenses of robot hardware and the severe cognitive bottlenecks of teleoperation.

### C. Robot-only Ablation Study

To address **Q2**, we conduct an ablation study where only **Stage 2** from Section III-C, training with 50 robot-only data for each task, is carried out on a fresh ACT (implemented through [10]) 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 **30** independent physical hardware trials per task. As shown in Figure 4, the policy yields overall success rates of **33.3%, 50%, 40%, 36.7%** respectively for **T1, T2, T3**, and **T4**. The cumulative scores across all 30 trials are **17.5, 20, 14.5**, and **18** respectively for the four tasks, demonstrating that our teleoperation pipeline yields a capable standalone baseline for multi-robot learning.

### D. DUET Performance

To address **Q3**, we execute **Stage 1** (Section III-C) 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 4, where each experiment is conducted 30 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 IV-B, 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 I. 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 IV-B. 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 I: **Data Collection Effort in Minutes.**

## V. Conclusion

We introduced a multi-human pretraining pipeline, leveraging collaborative human data as a prior to accelerate and enhance dual-robot visuomotor policies. Our evaluations demonstrate that policies pretrained on human data yield task performance that meets or exceeds baselines trained on robot data of comparable collection time. This underscores the immense potential of large-scale human data as a pretraining foundation for multi-robot design.

## References

[1] Y. Zhao, R. Ge, C. S. Wang, J. Ye, M. Xu, M. Li, S. Zakharov, B. Van Hoorick, V. C. Guizilini, L. Guibas *et al.*, "Duet: Dual-robot understanding via efficient teaching," *arXiv preprint arXiv:2606.20990*, 2026.

[2] 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.

[3] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng *et al.*, "Egomimic: Scaling imitation learning via egocentric video," in *2025 IEEE International Conference on Robotics and Automation*, 2025.

[4] X. Yang, D. Kukreja, D. Pinkus *et al.*, "Sam 3d body: Robust full-body human mesh recovery," *arXiv preprint arXiv:2602.15989*, 2026.

[5] 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.

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

[7] Z. Luo, Y. Yuan, T. Wang *et al.*, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," *Science Robotics*, 2026.

[8] 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.

[9] 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*, 2016.

[10] R. Cadene, S. Alibert, A. Soare *et al.*, "Lerobot: State-of-the-art machine learning for real-world robotics in pytorch," https://github.com/huggingface/lerobot, 2024.