EMMA: Scaling Mobile Manipulation via Egocentric Human Data

Lawrence Y. Zhu[1], Pranav Kuppili[1*], Ryan Punamiya[1*], Patcharapong Aphiwetsa[1†], Dhruv Patel[1†], Simar Kareer[1], Sehoon Ha[1], Danfei Xu[1]

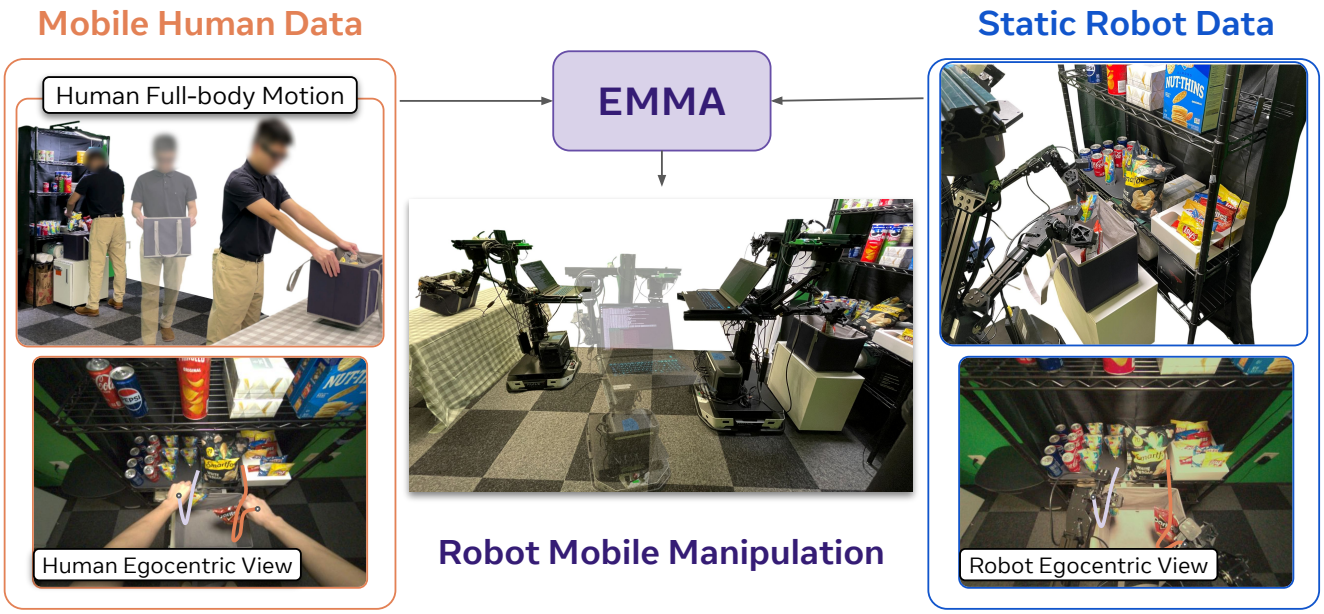

Fig. 1: EMMA learns mobile manipulation policies *without* collecting mobile manipulation teleoperation data. We achieve this through bridging embodiment kinematic gaps and unified co-training of mobile human data and static robot data.

***Abstract*—Scaling mobile manipulation imitation learning is bottlenecked by expensive mobile robot teleoperation. We present Egocentric Mobile MAnipulation (EMMA), an end-to-end framework training mobile manipulation policies from human mobile manipulation data with static robot data, sidestepping mobile teleoperation. To accomplish this, we co-train human full-body motion data with static robot data. In our experiments across three real-world tasks, EMMA demonstrates comparable performance to baselines trained on teleoperated mobile robot data (Mobile ALOHA), achieving higher or equivalent task performance in full task success. We find that EMMA is able to generalize to new spatial configurations and scenes, and we observe positive performance scaling as we increase the hours of human data, opening new avenues for scalable robotic learning in real-world environments. Details of this project can be found at https://ego-moma.github.io/.**

I. Introduction

Mobile manipulation has emerged as one of the most challenging problems in robotics due to the dual demands of navigation and manipulation. While recent advances in robot policy learning have demonstrated impressive capabilities in static manipulation, extending these successes to mobile scenarios introduces substantial challenges. The primary obstacle is data scarcity; current approaches tackling mobile manipulation rely on teleoperation frameworks akin to Mobile ALOHA [1] which face scalability limitations. This fundamental bottleneck limits dataset diversity and deployment robustness, particularly in unpredictable real-world settings where robots must navigate novel spatial configurations while maintaining manipulation precision.

Concurrently, recent work in cross-embodiment learning has explored human data as a scalable source of data for training end-to-end robot policies. Critically, human video data is cheap to collect, does not require a physical robot, and can be collected ergonomically via XR wearables [2], [3]. Human video data has been leveraged to train better visual representations [4], extract high-level object affordance [5], [6], [7], [8], or even as direct action supervision via co-training [2], [3]. However, these works have predominantly focused on table-top manipulation.

To this end, we introduce Egocentric Mobile Manipulation (EMMA), an end-to-end system to train mobile manipulation policies solely from *robot static-manipulation* and *human mobile-manipulation* data. While the robot data is captured via teleoperation, the human data is captured simply by wearing Project Aria glasses [9], thus avoiding costly mobile manipulation teleoperation. EMMA represents a step toward a new data paradigm in robot learning, where we imbue a robot with new skills by combining a set of robot teleoperation data with a more diverse pool of human demonstrations. In this work, our robot teleoperation data contains no mobile manipulation demonstrations, and we show that this skill can be effectively transferred from human data.

[1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30332, {lawrencezhu, pkuppili3, rpunamiya6, paphiwetsa3, dpatel756, skareer6, sehoonha, danfei}@gatech.edu

∗, † denotes equal contributions

EMMA presents a *full-stack framework* consisting of 1) an action retargeting pipeline which translates human mobile manipulation actions to a bimanual robot with a differential-drive base, 2) a unified architecture designed for co-training on heterogeneous human and robot data, and 3) an auxiliary phase identification mechanism that modulates between navigation and manipulation modes during inference. Therefore, it can switch control modes between navigation and manipulation phases based on predictions, preventing unintended base drift during manipulation and out-of-distribution arm movements during navigation.

We evaluate EMMA across three real-world mobile manipulation tasks: *table service*, *handover wine*, and *grocery shopping*. Our experiments reveal three key findings. First, EMMA achieves superior performance compared to baselines trained on teleoperated mobile robot data, showing that human demonstrations can replace costly mobile teleoperation. Second, we observe favorable scaling properties—each additional portion of human data yields greater performance than an equivalent portion of teleoperated robot data. Third, EMMA exhibits robust generalization performance, successfully transferring navigation and manipulation skills to novel environments seen only in human demonstrations, unlike teleoperation-based approaches that fail to adapt. These results open new avenues for scalable robotic learning in real-world environments.

II. Related Work

Behavior Cloning (BC). Behavior Cloning (BC) has emerged as an effective approach for robot learning, where policies are trained with direct supervised learning from expert demonstrations. Recent advances have shown remarkable results [10], [11], [12], [13], [14], [15], including the promise of building general-purpose policies by learning from large-scale datasets [11], [15], [16]. In particular, research around large multi-embodiment datasets such as Open-X [15] represents a significant milestone, demonstrating how models trained with diverse robot embodiments can acquire generalizable skills across tasks without task-specific engineering. However, most of these approaches focus on static manipulation tasks, with mobile manipulation remaining a significant challenge.

Learning for Mobile Manipulation. Building on successes in static manipulation, recent works have explored learning-based approaches for mobile manipulation that include skill primitives [17], [18], [19], reinforcement learning with decomposed action spaces [20], [21], [22], and whole-body control objectives [23], [24], [25]. Unlike these approaches, end-to-end imitation learning enables mapping raw pixel information to whole-body actions, showing promising results through large-scale training [26], [27], [28], [29], [11]. A critical challenge is to collect high-quality mobile manipulation demonstrations. Pioneering works [30], [31] have developed new teleoperation systems to facilitate data collection, including tethered leader-follower system [31], VR headsets [32], [33], motion-capture suits [34], [35], [36], smartphone-based control [37], kinesthetic teaching [38], and full-body teleoperation for humanoid robots [31] . However, despite these advances, collecting high-quality mobile manipulation data for diverse scenarios at scale remains a challenge. We propose to leverage egocentric human data (data collected with first-person perspective) captured by wearable devices as an alternative for scaling up imitation learning for mobile manipulation.

Robot Learning from Human Data. Recent work has focused on two complementary themes—leveraging human videos to bootstrap robot learning and finetuning with reinforcement learning for robust policies. In manipulation, co-training on paired egocentric human and robot demonstrations has been shown to boost skill performance [2], while zero- or few-shot transfer from human videos can be enabled by image inpainting and motion-track priors [39], [40]. Building on these, hierarchical planners extract latent action sequences from humans and distill them via retargeters into whole-body controllers [5], [31], [41]. In navigation, fusing imitation with RL—by turning keypoint matches into rewards or bootstrapping from behavior cloning—yields more reliable policies [42], [43], and inverse-dynamics models can pseudo-label passive egocentric video to distill intent-conditioned affordance subroutines [44]. We aim to train mobile manipulation policies from human data in a unified learning framework.

III. Hardware and Data Preliminaries

Egocentric Data Collection. In this work, we leverage a wearable smart glass Meta Project Aria [9] as our main data collection platform. Echoing prior work [2], we believe Aria glasses are ideal for capturing egocentric human data due to their ergonomic design and machine perception capability provided by the Machine Perception Service (MPS). Specifically, we leverage Aria glasses to capture both *exteroception* (wide-FOV egocentric RGB images) and *proprioception* (hand tracking and global localization) data in human mobile manipulation behaviors.

Low-cost Bimanual Mobile Manipulator. To effectively utilize egocentric human data for mobile manipulation, the robot hardware platform must resemble human sizes and kinematic workspaces. Drawing inspiration from the "Eve" robot platform introduced in EgoMimic [2], we develop a low-cost custom mobile manipulator that comprises of two 6-DoF ViperX 300s mounted in an identical inverted configuration on a height-adjustable rig. The rig is mounted on an AgileX TRACER differential drive AGV platform, which is capable of moving up to 2m/s. The full system stands a maximum height of 1.75m. Similar to Eve, we propose to leverage Aria glasses as the main egocentric perception sensor for the robot and mount it in a way that emulates the hand-eye configuration of a human adult. This mitigates the human-robot camera device gap and reduces the sensor-manipulator kinematic gap. Each arm is equipped with an Intel Realsense D405 on its wrist to facilitate precise near-range manipulation.

Human and Robot Data Streams. We collect distinct data streams from both human demonstrations ($\mathcal{D}_H$) and robot execution ($\mathcal{D}_R$). A key shared data stream is the egocentric RGB image I_{ego}, generated by the Aria glasses worn by the human or mounted on the robot. The robot provides RGB streams from its two wrist cameras, I_{wrist}. Proprioceptive data streams capture the state of each embodiment: For the human,

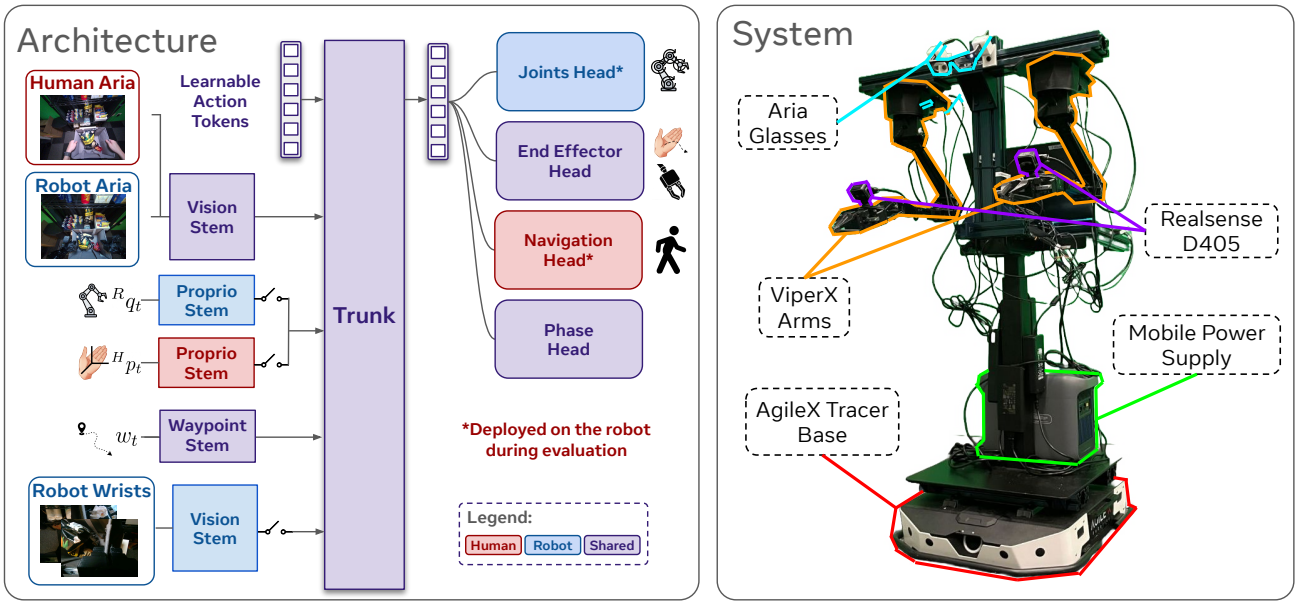

Fig. 2: Left: Architecture of joint human-robot policy learning framework. Our model processes heterogeneous human and robot data through *stems* and decodes them through various action *heads*. The navigation head is deployed on the robot during evaluation, demonstrating transfer without robot supervision. Right: Our custom low-cost bimanual mobile manipulator.

we leverage the Aria Machine Perception Service (MPS) [45] to estimate bimanual 3D hand poses $H_p \in SE(3) \times SE(3)$ and the 3D head pose $H_d \in SE(3)$ (in a SLAM-based world frame). For the robot, proprioception includes the state of its bimanual arms via joint positions $R_q \in \mathbb{R}^{2\times 7}$ (including gripper state) and the corresponding end-effector poses $R_p \in SE(3) \times SE(3)$.

For navigation, we process the human head pose H_d to extract 2D base pose representations. Specifically, we project the 3D head pose onto the ground plane to obtain $h^t_{\text{base}} = (x^t, y^t, \theta^t) \in SE(2)$, where (x^t, y^t) represents the 2D position and θ^t represents the yaw angle at time t. We maintain a displacement-based waypoint history $\mathcal{W}_t = \{h^{t-k_i}_{\text{base}}\}_{i=1}^{K_h}$, where K_h is the maximum number of historical waypoints. New waypoints are added when the displacement $\|h^t_{\text{base}} - h^{t-k_i}_{\text{base}}\|_2 \geq d_{\text{thresh}}$ (e.g., every 0.5m). This displacement-based sampling ensures consistent spatial resolution regardless of movement speed. For policy learning, these waypoints are transformed into the current egocentric frame as $\tilde{\mathcal{W}}_t = \{T_{ego}^{-1} \cdot h^{t-k_i}_{\text{base}}\}_{i=1}^{K_h}$, where T_{ego} is the transformation from world to current egocentric coordinates. These egocentric waypoints provide speed-invariant spatial context for navigation action prediction.

IV. EMMA: System and Algorithm

EMMA is a scalable *full-stack system*, which enables (1) *direct transfer* of navigation skills from egocentric human data to a differential-drive mobile manipulator and (2) *scaling up* full mobile manipulation policy performance by co-training on both human mobile manipulation data and robot static manipulation data.

A. Data Retargeting and Alignment

A fundamental challenge in leveraging human data ($\mathcal{D}_H$) for robot learning lies in the significant embodiment gap, impacting both navigation and manipulation. Humans navigate omnidirectionally with decoupled head-gaze, whereas our robot uses a differential-drive base with kinematically constrained movements. Similarly, human hand motions (H_p) captured egocentrically differ kinematically and distributionally from robot end-effector motions (R_p). To enable effective knowledge transfer and co-training in an end-to-end imitation learning setting (Sec. IV-B), we introduce distinct retargeting and alignment strategies for navigation and manipulation.

Bridging Navigation Kinematic Gap. The primary focus of our retargeting efforts is translating human navigation trajectories into commands suitable for our differential-drive robot (Fig. 3). During human demonstrations, we extract 2D base poses $h^t_{\text{base}} = (x^t, y^t, \theta^t)$ by projecting the 3D head pose H^t_d onto the ground plane. These poses form navigation waypoints that capture the human's intended path.

However, directly mapping this sequence of waypoints to robot base commands is ill-posed due to differential-drive constraints (the robot can only move in straight lines and circular arcs) and the fixed alignment between the robot's torso-mounted Aria sensor and its heading. To overcome this, we formulate an optimization problem: given a sequence of desired waypoints $\{h^k_{\text{base}} = (x^d_k, y^d_k, \theta^d_k)\}_{k=1}^{K}$ extracted from human trajectory, find velocity commands $\mathbf{z} = [(v_1, \omega_1), ..., (v_K, \omega_K)]$ that minimize:

$$\min_{\mathbf{z}} \sum_{k=1}^{K} \Big[\lambda_{\text{pos}} \|p_k(\mathbf{z}) - p^d_k\|_2^2 + \lambda_{\text{yaw}} \text{wrap}(\theta_k(\mathbf{z}) - \theta^d_k)^2 + \lambda_{\text{smooth}} \left((v_k - v_{k-1})^2 + (\omega_k - \omega_{k-1})^2 \right) \Big] \quad (1)$$

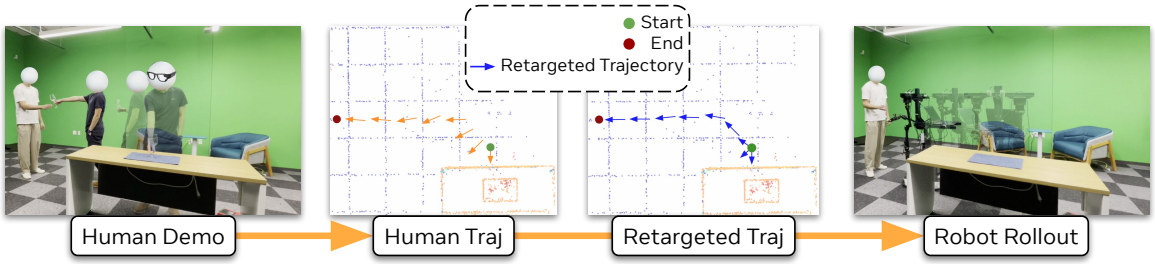

Fig. 3: Given a 3D pose trajectory from a human demonstration, we optimize Eq. 1 to compute a feasible 2D trajectory for a differential drive robot. This resulting trajectory can be directly executed by the robot or used as input for policy learning.

subject to the differential-drive dynamics:

$$x_{k+1} = x_k + v_k \cos(\theta_k)\Delta t \tag{2}$$

$$y_{k+1} = y_k + v_k \sin(\theta_k)\Delta t \tag{3}$$

$$\theta_{k+1} = \theta_k + \omega_k \Delta t \tag{4}$$

and constraints $v_{\min} \le v_k \le v_{\max}$, $\omega_{\min} \le \omega_k \le \omega_{\max}$.

Here, $p_k(\mathbf{z}) = (x_k, y_k)$ represents the robot position at step k resulting from applying the velocity sequence $\mathbf{z}$, $p_k^d = (x_k^d, y_k^d)$ is the desired position from the human waypoint, and wrap$(\cdot)$ ensures angular differences are in $[-\pi, \pi]$. The weights λ_{pos}, λ_{yaw}, and λ_{smooth} balance position tracking, heading alignment, and velocity smoothness respectively. This constrained optimization yields smooth, kinematically feasible trajectories that approximate human navigation patterns while respecting differential-drive constraints.

Aligning Manipulation Action Data. For manipulation, we address mismatches between human hand data (H_p) and robot end-effector data (R_p). Inspired by prior work like EgoMimic [2], we first unify coordinate frames by transforming all upper-body action chunks (both human and robot) into the reference frame of the camera *at the time of observation*, using SLAM estimates for human data and hand-eye calibration for robot data. This makes predictions relative to the current view. Second, acknowledging persistent distributional gaps due to biomechanics and sensors, we apply Z-score normalization independently to the transformed pose and action data within each source (D_H and D_R), using their respective dataset statistics.

B. Human and Robot Data Co-training

The goal of our system is to transfer knowledge from egocentric human mobile manipulation data (D_H) to scale policy performance, while still leveraging limited static robot manipulation data (D_R) for precise manipulation steps. The re-targeting and action alignment steps detailed in Sec. IV-A allow us to treat human and robot data as equal parts in a continuous spectrum of embodied data sources. But large domain gaps exist in the two data sources, in both sensing modalities and distributions. Thus, effectively learning from these data necessitates a unified learning framework capable of processing *heterogeneous* data sources. Inspired by recent works in cross-embodiment policy learning [46], [47], [48], we design an architecture based on a decoder-only Transformer with modality-specific input *stems*, a shared *trunk*, and multiple action and auxiliary output *heads* (Fig. 2).

Stems. Stems are shallow networks that encode raw observations from different modalities into a sequence of fixed-dimension tokens. Crucially, we employ a *shared* vision stem for processing the main egocentric RGB images (I_{ego}) from the Aria glasses (human and robot) to enforce visual feature alignment. Separate stems handle inputs from the robot's wrist cameras (I_{wrist}), which does not exist in human data.

Trunk. The Trunk is a standard multi-layer decoder-only transformer that processes the concatenated token sequences from all active stems. A sequence of M learnable tokens, representing the action chunk length, is prepended to the input sequence constructed from all of the stems.

Heads. The heads are shallow MLPs that map the first M output tokens from the trunk to the respective action spaces or auxiliary predictions. We define four heads for predicting robot bimanual joint actions ($\mathbb{R}^{K\times 14}$), human cartesian end-effector actions ($\mathbb{R}^{K\times 3}$), robot base navigation actions ($(x, y, \omega) \in \mathbb{R}^{K\times 3}$), and, as an auxiliary output, the predicted task phase ($p \in \{0, 1\}^K$, see Sec. IV-C).

Co-training from heterogeneous sources. The model is trained jointly on batches drawn from two sources: (1) the collected human mobile manipulation demonstrations D_H (with navigation actions processed by the retargeting module described in Sec. IV-A) and (2) static robot manipulation demonstrations D_R (e.g., collected via teleoperation). When processing a human data batch, the human proprioception, shared ego vision stem, human manipulation action head, navigation action head and the shared phase prediction head are active. For a robot batch, the robot proprioception, wrist image stem, shared ego vision stem, and robot manipulation action head are active. The shared trunk and ego vision stem is updated by all sources and modalities, forcing them learn versatile representations. The navigation head primarily learns from the retargeted human data, transferring human navigation strategies to the robot. The complete architecture is illustrated in Fig. 2

C. Auxiliary Phase Identification and Control Modulation

Mobile manipulation tasks naturally alternate between navigation and manipulation phases, requiring different control strategies. We introduce an unsupervised phase identification mechanism that automatically segments demonstrations and modulates control during deployment.

Phase Detection. We identify phases based on motion dynamics. For each frame, we compute the ratio of hand velocity to head velocity. Manipulation phases exhibit high hand-to-head velocity ratios ($> \tau_{ratio}$) with low absolute head velocity ($< \tau_{head}$), indicating dexterous hand motion while stationary. We filter candidate segments shorter than $\tau_{duration}$ frames to remove transient gestures.

Algorithm 1 Unsupervised Phase Identification

Input: Head poses $\mathbf{p}_{head}$, hand poses $\mathbf{p}_{hand}$
Output: Phase labels $\phi \in \{0, 1\}^N$
$v_{head}, v_{hand} \leftarrow \|\Delta \mathbf{p}_{head}\|/\Delta t, \|\Delta \mathbf{p}_{hand}\|/\Delta t$
$r \leftarrow v_{hand}/(v_{head} + \epsilon)$
$\mathcal{M} \leftarrow \{i : r_i > \tau_{ratio} \wedge v_{head,i} < \tau_{head}\}$
Fit GMM(K) on $\{\mathbf{p}_{head,i} : i \in \mathcal{M}\}$
$\phi_i \leftarrow \text{GMM.pdf}(\mathbf{p}_{head,i}) < \tau_{pdf}$

To spatially localize K manipulation zones, we fit a Gaussian Mixture Model (GMM) with K components to head positions during these high-ratio periods. Each frame is classified as manipulation (phase 0) if its probability density under the GMM exceeds τ_{pdf}, otherwise navigation (phase 1). GMM provides direct probability density estimates needed for phase classification without requiring labeled training data, while being computationally efficient and interpretable for spatial clustering.

Phase-Aware Control: During deployment, predicted phases modulate the navigation action chunk to prevent unintended movements:

- **Manipulation phases:** Navigation actions interpolate from zero to the first future navigation waypoint.
- **Navigation phases:** Any future manipulation-phase positions are replaced with the current navigation endpoint.

This phase-aware modulation removes the base action noise introduced from head movements during manipulation and ensures smooth switch between the two phases.

V. Experiments

We aim to validate three key hypotheses. **H1:** EMMA can achieve performance comparable to systems trained on teleoperated mobile manipulation data. **H2:** Key design decisions of EMMA improve downstream task performance and robustness. **H3:** Given an initial amount of static robot manipulation data, it is more valuable to collect addition human mobile manipulation over mobile robot teleoperation data. We evaluate these hypotheses through three long horizon mobile manipulation tasks.

Table Service. Two tables are set 2m apart. The kitchen table has an oven with four croissants, a plate, and wrapped utensils. The dining table has a mat with a wine glass in the corner. The robot picks up the utensils and navigates to the dining table, placing them on the left side. It then returns to the kitchen table, picks and places a croissant onto the plate, and navigates back to place the plate in the center of the dining mat, avoiding the wine glass.

Handover Wine. The robot picks up a wine glass randomly placed within a 30cm×45cm area on a table. It turns right and navigates toward a human standing in a 3m×3m area. At a safe range, the robot hands the wine glass to the human's right hand, testing precise navigation and state coverage transfer from human data.

Grocery Shopping. The robot faces a grocery shelf and simultaneously grabs a juice pouch from the left and a chip bag from the right, placing them into a shopping bag. It then uses both arms to pick a large popcorn bag from the center shelf and add it to the bag. Finally, the robot lifts the shopping bag and navigates to a table behind it, testing long-horizon bimanual manipulation and navigation.

Baselines. We implement **Mobile ALOHA** [1] as our primary baseline to compare against teleoperated mobile robot data. Critically, it exemplifies the exact data collection paradigm—expensive mobile teleoperation—that EMMA aims to replace with human egocentric data. To ensure fair comparison, we modify it to the same HPT [47] backbone used in EMMA, isolating the comparison to data sources (human vs. teleoperation) rather than architectural differences. This baseline receives identical input modalities (egocentric RGB, wrist cameras, proprioception) and outputs robot joint actions $R_q \in \mathbb{R}^{14}$ plus base velocities $(v, \omega) \in \mathbb{R}^2$ recorded from the AgileX Tracer wheel encoders.

In addition, we evaluate two ablation baselines. **EMMA w/o action retargeting** replaces our optimization-based retargeting (Eq. 1) with raw human navigation actions, testing whether kinematic alignment is necessary for successful transfer. **EMMA w/o phase identification** removes the phase identification mechanism that modulates control during deployment, allowing both arm and base to move simultaneously throughout execution.

Evaluation protocols. To ensure statistical significance of the results, We conduct 50 trials per task/model variant configuration, and record error bars with 95% confidence interval, while recording both success metrics and failure modes for qualitative analysis (Fig. 8). Each trial has a maximum duration of 2 minutes. For method comparison across all baseline and ablation experiments, we ensure identical environmental conditions and object positions. Alltogether, we conducted *950 mobile manipulation rollout* evaluations to produce the results in this paper.

A. Main Results

EMMA achieves favorable performance compared to systems trained on teleoperated mobile robot data (H1).

EMMA consistently outperforms teleoperation-based baselines across a range of mobile-manipulation tasks by leveraging egocentric human demonstrations (Fig. 4). In the *Handover Wine* task, with the same 100 demos of static robot data, replacing one hour of teleoperated mobile robot data with one hour of human mobile manipulation yields an 82% success rate—a 30% increase over teleoperated robot data, which had a 52% success rate. This improvement demonstrates the benefit of learning from egocentric human-human interaction. Leveraging egocentric human mobile manipulation data has also shown improvement in bimanual tasks found in the *Grocery Shopping* task. Lastly, using egocentric human mobile manipulation data has shown comparable performance to teleoperated

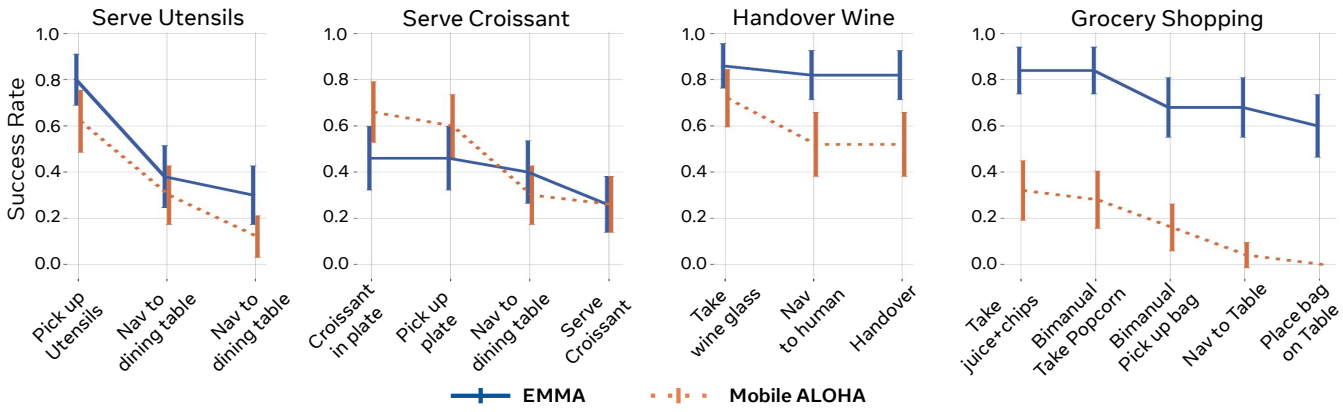

Fig. 4: Cumulative success rates across subtasks for three mobile manipulation tasks. EMMA (blue), trained without mobile teleoperation data, significantly outperforms Mobile ALOHA (orange) on *Grocery Shopping* and *Handover Wine* tasks ($p < 0.05$). *Table Service* variants show comparable performance. Error bars represent 95% confidence intervals from 50 trials.

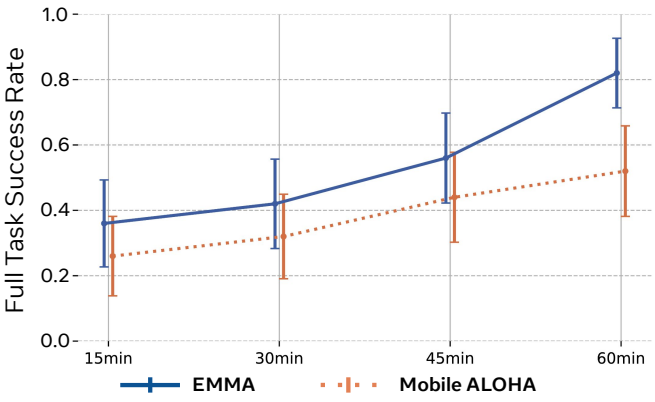

Fig. 5: For *Handover Wine* task, starting with a fixed amount of static manipulation data, we show that adding more human fullbody motion data for EMMA (blue) yields to greater performance gains compared to adding mobile robot teleoperation data collected under an equivalent amount of time for Mobile ALOHA (orange). The performance gap expands from 10% to 30% as data increases from 15 to 60 minutes.

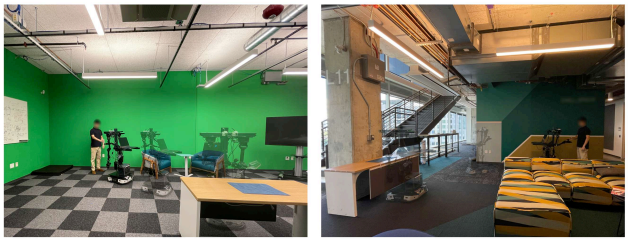

Fig. 6: EMMA co-trains static robot data in lab (left) with human demonstrations in a novel scene (right) to enable robot deployment in previously unseen environments.

mobile robot data on long-horizon *Table Service* task, a task that requires long-distance point-to-point navigation, where robot-teleoperated methods benefit from zero perspective gap. Taken together, these results show that egocentric human mobile manipulation data matches or even improves the value

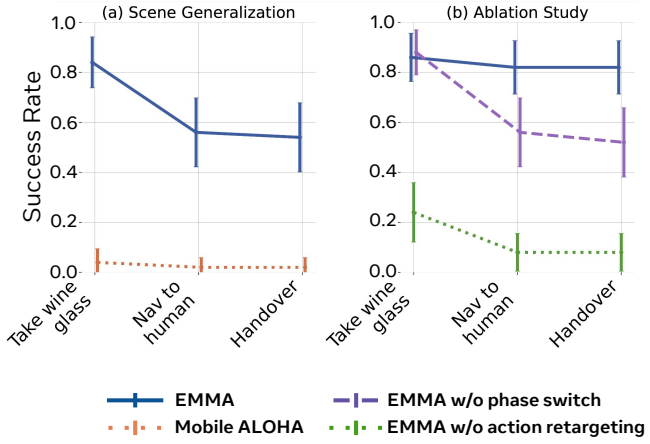

Fig. 7: **(a)** EMMA generates to an unseen scene with 54% full task success rate. (b) **Ablation study** on the *Handover Wine* task. Removing either retargeting or phase switch causes significant performance drop.

of equivalent robot teleoperation data and delivers safer, more reliable, and more scalable policies across diverse mobile-manipulation challenges.

EMMA generalizes to unseen scenes. We evaluate scene generalization by testing EMMA in a novel environment seen only through human demonstrations. Specifically, we collect 30 minutes of human demonstrations in a new spatial layout where the wine recipient stands randomly within a larger 5m $\times$ 2m area (Fig. 6). Despite never seeing robot data from this environment, EMMA achieves 54% success rate (Fig. 7 (a)), demonstrating two key capabilities: (1) navigation behaviors learned from human data naturally adapt to the expanded spatial configuration, and (2) manipulation skills remain robust to visual domain shifts from the new environment. In contrast, Mobile ALOHA—trained exclusively on teleoperated data from the original environment is unable to complete the initial grasp due to the environmental changes (Fig. 8 (j)). This experiment validates that human data provides superior generalization through its inherent diversity compared to lab-constrained teleoperation.

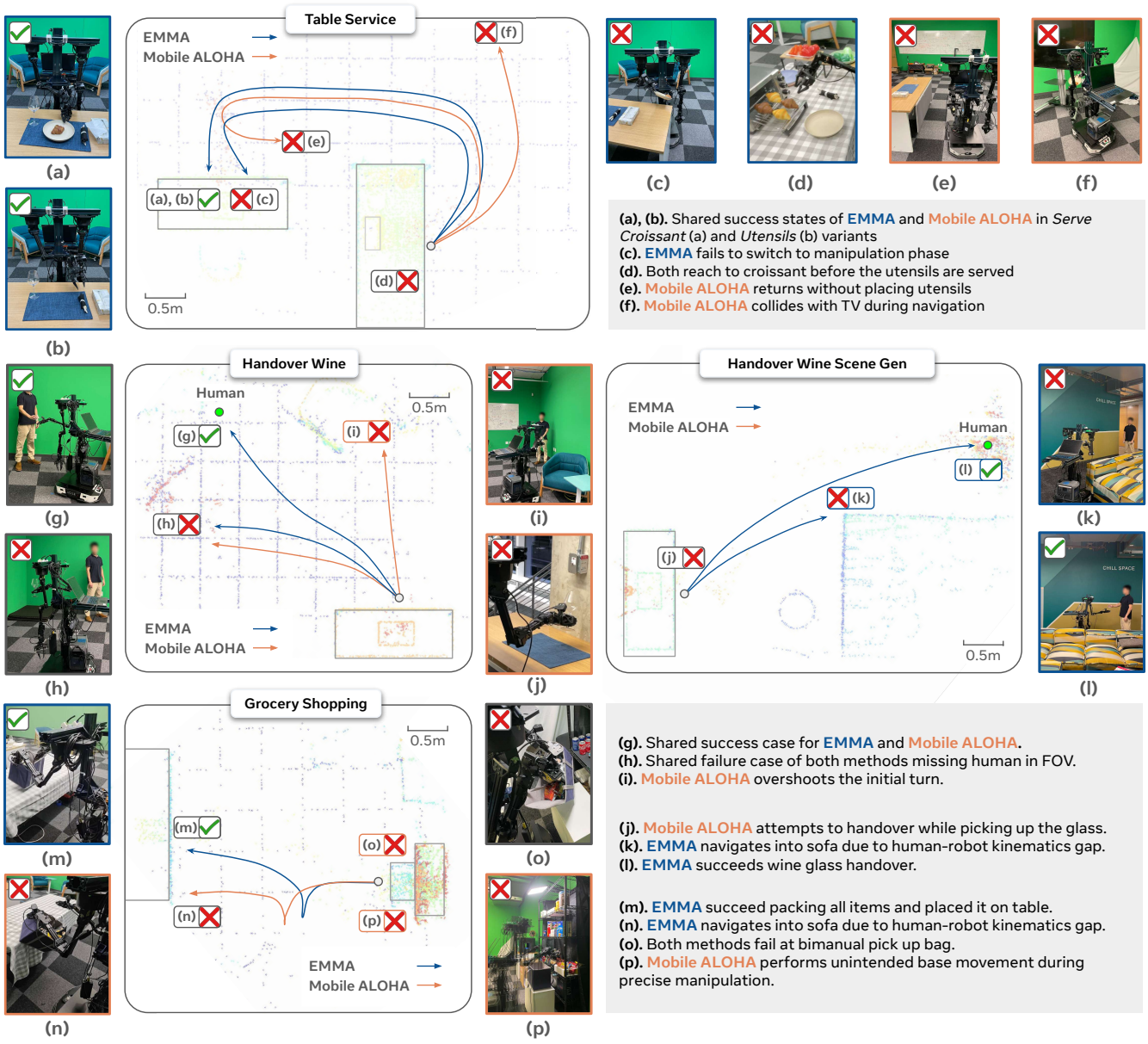

Fig. 8: Qualitative comparison of EMMA and Mobile ALOHA across three mobile manipulation tasks, showing representative success and failure modes.

Ablation Studies (H2). Our ablation experiments on the *Handover Wine* task (Fig. 7 (b)) reveal the critical importance of both key components. Without kinematic retargeting, the success rate drops by 30%, as the robot frequently loses track of the human recipient when executing kinematically infeasible trajectories. The phase identification mechanism proves equally crucial—its removal causes complete task failure due to unintended base movements during grasping and unnecessary arm motions during navigation, resulting in collisions and dropped objects. These results validate that both human-robot embodiment alignment and learned phase modulation are essential for successful human-to-robot skill transfer in mobile manipulation.

Human data scales mobile manipulation performance more efficiently compared to teleoperated robot data (H3). Keeping one hour of static robot manipulation data fixed, EMMA's success rate climbs steadily from 0.36 to 0.82 as we increase human mobile manipulation data from 15 minutes to one hour. Mobile ALOHA also improves—from 0.26 to 0.52—when its robot teleoperation data grows from 15 minutes to one hour, but remains consistently below EMMA. This demonstrates that additional human mobile manipulation demonstrations yield substantially greater returns than equivalent increases in mobile robot teleoperation data. Qualitatively EMMA shows better robustness to target pose, fewer collisions with the environment and greater overall success. This is summarized in Fig. 8

VI. Conclusion

In conclusion, we presented EMMA, a novel framework that enables mobile manipulation without requiring expensive mobile teleoperation. We find that EMMA outperforms Mobile ALOHA style teleoperation with equivalent data collection time and demonstrates superior scaling properties. Further, we ablate our key design decisions and observe that both the motion retargeter and the phase identification module are integral to downstream performance. Overall, we have demonstrated the possibility of scaling robot mobile manipulation performance with egocentric human data. We hope this result spurs new research and informs future researchers of potential opportunities and challenges.

VII. Acknowledgment

This work was partially supported by the Samsung Research America LEAP-U Program, the Industrial Technology Innovation Program (P0028404) of the Ministry of Industry, Trade and Energy of Korea, and a research gift from Meta Platforms, Inc.

References

[1] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conf. Robot Learn. (CoRL)*, 2024.

[2] S. Kareer *et al.*, "Egomimic: Scaling imitation learning via egocentric video," 2024.

[3] R. Qiu *et al.*, "Humanoid policy ˜human policy," *arXiv preprint arXiv:2503.13441*, 2025.

[4] S. Nair *et al.*, "R3M: A universal visual representation for robot manipulation," 2022.

[5] C. Wang *et al.*, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[6] M. Xu *et al.*, "Flow as the cross-domain manipulation interface," 2024.

[7] C. Wen *et al.*, "Any-point trajectory modeling for policy learning," 2024.

[8] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," 2024.

[9] J. Engel *et al.*, "Project aria: A new tool for egocentric multi-modal AI research," 2023.

[10] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *arXiv preprint arXiv:2307.15818*, 2023.

[11] M. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[12] K. Black *et al.*, "π0: A vision-language-action flow model for general robot control," *URL https://arxiv.org/abs/2410.24164*, 2024.

[13] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *Int. J. Robot. Res.*, 2024.

[14] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[15] A. O'Neill *et al.*, "Open x-embodiment: Robotic learning datasets and RT-X models," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 6892–6903.

[16] A. Khazatsky *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[17] C. Sun *et al.*, "Fully autonomous real-world reinforcement learning with applications to mobile manipulation," in *Conf. Robot Learn.*, 2022, pp. 308–319.

[18] B. Wu, R. Martin-Martin, and L. Fei-Fei, "M-ember: Tackling long-horizon mobile manipulation via factorized domain transfer," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 11 690–11 697.

[19] J. Wu *et al.*, "Tidybot: Personalized robot assistance with large language models," *Auton. Robots*, vol. 47, no. 8, pp. 1087–1102, 2023.

[20] J. Gu *et al.*, "Multi-skill mobile manipulation for object rearrangement," *arXiv preprint arXiv:2209.02778*, 2022.

[21] N. Yokoyama *et al.*, "ASC: Adaptive skill coordination for robotic mobile manipulation," *IEEE Robot. Autom. Lett.*, vol. 9, no. 1, pp. 779–786, 2023.

[22] F. Xia *et al.*, "Relmogen: Integrating motion generation in reinforcement learning for mobile manipulation," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 4583–4590.

[23] Z. Fu, X. Cheng, and D. Pathak, "Deep whole-body control: learning a unified policy for manipulation and locomotion," in *Conf. Robot Learn.*, 2023, pp. 138–149.

[24] R. Yang *et al.*, "Harmonic mobile manipulation," in *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2024, pp. 3658–3665.

[25] J. Hu, P. Stone, and R. Martín-Martín, "Causal policy gradient for whole-body mobile manipulation," *arXiv preprint arXiv:2305.04866*, 2023.

[26] M. Ahn *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[27] A. Brohan *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[28] N. M. M. Shafiullah *et al.*, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023.

[29] Physical Intelligence *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," 2025.

[30] Y. Jiang *et al.*, "BEHAVIOR robot suite: Streamlining real-world whole-body manipulation for everyday household activities," *arXiv preprint arXiv:2503.05652*, 2025.

[31] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," in *Conf. Robot Learn. (CoRL)*, 2024.

[32] S. Chen *et al.*, "ARCap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback," 2024.

[33] C. Lu *et al.*, "Mobile-television: Predictive motion priors for humanoid whole-body control," 2025.

[34] R. Cisneros *et al.*, "Team JANUS humanoid avatar: A cybernetic avatar to embody human telepresence," in *Toward Robot Avatars: Perspectives on the ANA Avatar XPRIZE Competition, RSS Workshop*, vol. 3, 2022.

[35] S. Dafarra *et al.*, "iCub3 avatar system: Enabling remote fully immersive embodiment of humanoid robots," *Sci. Robot.*, vol. 9, no. 86, p. eadh3834, 2024.

[36] K. Darvish *et al.*, "Whole-body geometric retargeting for humanoid robots," in *IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, 2019, pp. 679–686.

[37] J. Wong *et al.*, "Error-aware imitation learning from teleoperation data for mobile manipulation," in *Proc. 5th Conf. Robot Learn.*, ser. Proc. Mach. Learn. Res., vol. 164, 2022, pp. 1367–1378.

[38] T. Yang *et al.*, "MOMA-Force: Visual-force imitation for real-world mobile manipulation," in *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2023, pp. 6847–6852.

[39] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," 2025.

[40] J. Ren *et al.*, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," *arXiv preprint arXiv:2501.06994*, 2025.

[41] Y. Yan, E. V. Mascaro, and D. Lee, "Imitationnet: Unsupervised human-to-robot motion retargeting via shared latent space," in *IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, 2023, pp. 1–8.

[42] H. Karnan, G. Warnell, X. Xiao, and P. Stone, "Voila: Visual-observation-only imitation learning for autonomous navigation," in *Int. Conf. Robot. Autom. (ICRA)*, 2022, pp. 2497–2503.

[43] R. Ramrakhya, D. Batra, E. Wijmans, and A. Das, "PIRLNav: Pretraining with imitation and RL finetuning for ObjectNav," in *CVPR*, 2023.

[44] A. Kumar, S. Gupta, and J. Malik, "Learning navigation subroutines from egocentric videos," 2019.

[45] Meta Reality Labs Research, "Machine perception services (MPS)," 2024, [Online]. Available: https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps.

[46] S. Liu *et al.*, "RDT-1B: A diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[47] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers," in *NeurIPS*, 2024.

[48] Octo Model Team *et al.*, "Octo: An open-source generalist robot policy," in *Proc. Robot.: Sci. Syst.*, 2024.

APPENDIX

A. Problem Formulation Details

1) Bi-source Imitation Learning Problem: We formalize our mobile manipulation learning problem as imitation learning from two heterogeneous data sources: human mobile manipulation demonstrations D_H and static robot manipulation demonstrations D_R. The observation spaces are:

- $\mathcal{O}^H$: Human observation space containing egocentric RGB images I_{ego}, hand poses $H_p \in SE(3) \times SE(3)$ in the egocentric RGB camera frame, and 3D head pose $H_d \in SE(3)$.
- $\mathcal{O}^R$: Robot observation space containing egocentric RGB I_{ego}, wrist camera images I_{wrist}, joint positions $R_q \in \mathbb{R}^{14}$, and end-effector poses $R_p \in SE(3) \times SE(3)$.

The shared egocentric view I_{ego} serves as the primary linking modality between embodiments. During deployment, we assume:

- **Manipulation phase:** $\mathcal{O}_{\text{deploy}} \subseteq \mathcal{O}^R \cup \mathcal{O}^H$
- **Navigation phase:** $\mathcal{O}_{\text{deploy}} \subset \mathcal{O}^H$

This assumption enables direct transfer of navigation knowledge purely from human data while leveraging both sources for manipulation.

B. Phase-Conditioned Policy

The policy operates under implicit phase conditioning:

$$\pi(a_t \mid o_t, \phi_t), \tag{5}$$

where $\phi_t \in \{0, 1\}$ indicates manipulation (0) or navigation (1). Phase identification uses unsupervised clustering with parameters:

- Velocity ratio threshold: $\tau_{\text{ratio}} = 2.0$
- Head velocity threshold: $\tau_{\text{head}} = 0.4$ m/s
- Minimum phase duration: $\tau_{\text{duration}} = 30$ frames (1 s at 30 fps)
- GMM components: $K = 2$ for spatial manipulation-zone identification
- Probability density threshold: $\tau_{\text{pdf}} = 10^{-3}$

C. Differential-Drive Retargeting Constraints

The optimization problem (Equations 1–4 in the main text) employs the following bounds and weights:

- Velocity constraints: $v \in [-1.0, 1.0]$ m/s, $\omega \in [-\pi, \pi]$ rad/s
- Optimization weights: $\lambda_{\text{pos}} = 32.0$, $\lambda_{\text{yaw}} = 2.0$, $\lambda_{\text{smooth}} = 1.0$
- Integration timestep: $\Delta t = 0.02 \times \text{STEP}$ seconds
- Navigation sampling step: $\text{STEP}_{\text{nav}} = 8$ frames in all trained models.

The retargeting ensures kinematic feasibility while minimizing deviation from human trajectories through constrained optimization.

Modality	Horizon	Sampling Step	Interpolated Length
Navigation	10	8 frames	100
Manipulation	10	4 frames	100

D. Action Chunking and Temporal Alignment

Actions are sampled at different temporal resolutions and then unified:

We interpolate to a fixed chunk length of 100 to ensure consistent action representation across modalities. Navigation actions undergo phase-aware filtering in which manipulation-phase waypoints are replaced with interpolated trajectories to the first navigation waypoint.

E. Coordinate Frame Alignment

All action predictions are expressed in the egocentric camera frame at observation time:

- **Human hand poses:** Transformed using the SLAM-estimated camera pose T_{ego}^{-1}.
- **Robot end-effector:** Transformed using the calibrated hand–eye transformation.
- **Waypoint history:** Maintained as $\tilde{W}_t = \{ T_{\text{ego}}^{-1} \cdot h_{\text{base}}^{t-k_i} \}_{i=1}^{K_h}$ with displacement threshold $d_{\text{thresh}} = 0.25\,\text{m}$.

Z-score normalization is applied independently per data source (D_H, D_R) to handle distributional differences.

F. Data Quality Constraints

The framework enforces several quality constraints:

- **Hand-tracking confidence filtering:** Frames with confidence < 0 are excluded, allowing data collectors to temporarily lift the glasses to reset a scene without corrupting supervision.
- **Waypoint sparsity:** New waypoints are added only when $\|h_{\text{base}}^{t} - h_{\text{base}}^{t-k}\|_2 \geq d_{\text{thresh}}$, enforcing spatial sampling at $\geq d_{\text{thresh}}$ increments.
- **Phase-transition smoothing:** Segments shorter than τ_{duration} are filtered.

G. Temporal Consistency Assumption

Phase transitions are assumed sparse, with manipulation zones spatially localized. This assumption—enforced via minimum-duration constraints and GMM-based spatial clustering—enables reliable phase prediction and prevents oscillatory switching during deployment.

H. Network Architecture

a) Architecture: Each input modality is encoded with a dedicated stem (e.g. a ResNet-18 for images and an MLP for low-dimensional state). Each stem produces a feature sequence that is tokenized via cross-attention using a small bank of learnable queries, yielding a fixed number of latent tokens per modality. Sinusoidal positional embeddings are added (i) within each modality stream prior to tokenization and (ii) again after concatenating all modality tokens. We then append a learnable bank of action tokens whose count matches the action horizon, and process the full sequence with a Transformer trunk (Fig. 9).

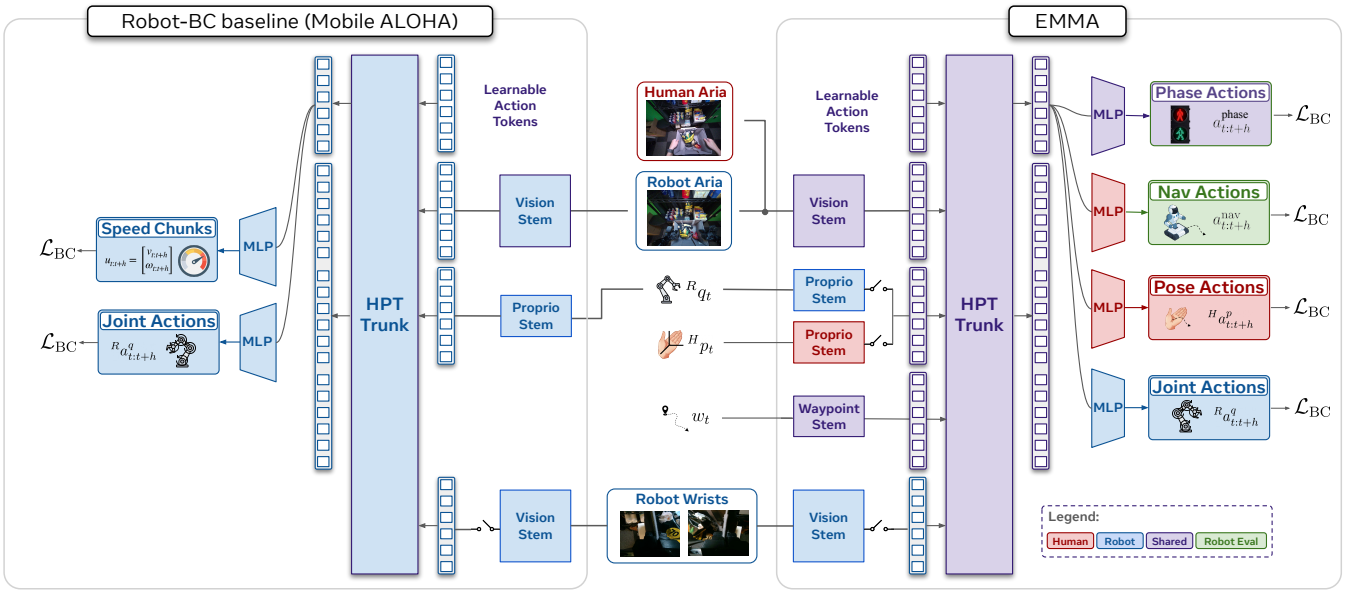

Fig. 9: Detailed network architecture comparison of Mobile ALOHA (left) and EMMA (right). We supervise all heads with Smooth-ℓ_1 (β=0.05) + MSE loss.

TABLE I: Hyperparameters for EMMA vs. Mobile ALOHA.

Item	**EMMA**	**Mobile ALOHA**
Input resolution $H \times W \times C$	$480 \times 640 \times 3$	$480 \times 640 \times 3$
Action chunk length k	100	100
Image encoder	ResNet-18 (ImageNet init.)	ResNet-18 (ImageNet init.)
Modality embedding dim d_{proj}	256	256
Latent queries per modality L	16	16
Stem cross-attn heads D_{stem}	8	8
Stem per-head dim d_{attn}	64	64
Transformer trunk hidden dim d	256	256
Transformer trunk heads D_{trunk}	8	8
Transformer trunk layers N_{trunk}	16	16
Learnable action queries	100	100
Positional encoding	Sinusoidal (per-stream & global)	Sinusoidal (per-stream & global)
Prediction head	MLP (hidden width 512, SiLU, LayerNorm)	MLP (hidden width 512, SiLU, LayerNorm)
Supervision loss	Smooth-ℓ_1 (β=0.05) + MSE	Smooth-ℓ_1 (β=0.05) + MSE
Nav. upsampling	$10 \rightarrow 100$ (linear x, y; atan2 yaw)	—
Action dims d_a	7 or 14 (robot joints) + 6 or 12 (human ee pose) + 3 (nav waypoints)	7 or 14 (robot joints) + 2 (base velocities)

b) Training targets and loss: All targets (e.g. manipulation trajectories, navigation waypoints, phase) are supervised directly with a smooth-ℓ_1 objective. When a shared head is active it is supervised with the corresponding shared target; auxiliary heads (when present) are supervised on their respective labels using the same backbone features.

I. Baseline Details

As discussed in Sec. V of the main text, we compare EMMA with Mobile ALOHA by varying the source of mobile-manipulation supervision. Like EMMA, Mobile ALOHA co-trains static robot manipulation data (teleoperation) with an additional source of mobile data (Fig. 9). However, Mobile ALOHA replaces human full-body data with mobile teleoperation, where the robot's base linear and angular velocities are recorded at 50 Hz using wheel encoders on the AgileX Tracer platform. For a fair comparison, the stem and trunk architectures remain identical to EMMA, while the prediction head is extended to output a chunk of future base velocities (Table I).

To maximize data collection efficiency, two teleoperators are employed: one controls the robot arms using a leader–follower system physically connected to the robot laptop via USB extension cables, while the other simultaneously drives the mobile base using a remote controller.