EGOX: EGOCENTRIC CROSS-EMBODIMENT MANIPULATION WITH EMBODIMENT DREAMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling robot manipulation requires learning from demonstrations collected across humans and heterogeneous robots. These sources cover diverse objects, scenes, tasks, and interaction strategies, but gaps in sensing, kinematics, motion patterns, and action spaces make it difficult to train a shared policy effectively. We introduce EgoX, an egocentric cross-embodiment framework for robot manipulation. Its central design is to share manipulation structure at the level of semantically corresponding modalities, while preserving embodiment-specific sensing and control interfaces. EgoX stores human and robot demonstrations as unified egocentric observation-action trajectories, pretrains a Modularized Cross-Embodiment Transformer with Embodiment Dreaming (MXT+) on the mixture, and fine-tunes it on target-embodiment demonstrations. MXT+ combines embodiment-specific modality tokenizers and action experts with a shared Transformer trunk. An auxiliary embodiment dreaming objective predicts future embodiment latents from egocentric visual and proprioceptive features, using exponential-moving-average copies of the modality encoders; this training-only branch encourages the trunk to encode embodiment-specific temporal structure without adding deployment-time computation. We evaluate EgoX on five manipulation tasks across five real robots (wheeled humanoid, legged humanoid, quadrupedal manipulator, two tabletop manipulation platforms) and one simulated humanoid, using human demonstrations as an additional pretraining source. Across in-distribution and out-of-distribution settings, cross-embodiment pretraining improves downstream manipulation, broader embodiment mixtures improve transfer, and embodiment dreaming provides additional gains on contact-rich tool-use tasks. Anonymous project website: `https://egox-embodiment.github.io`.

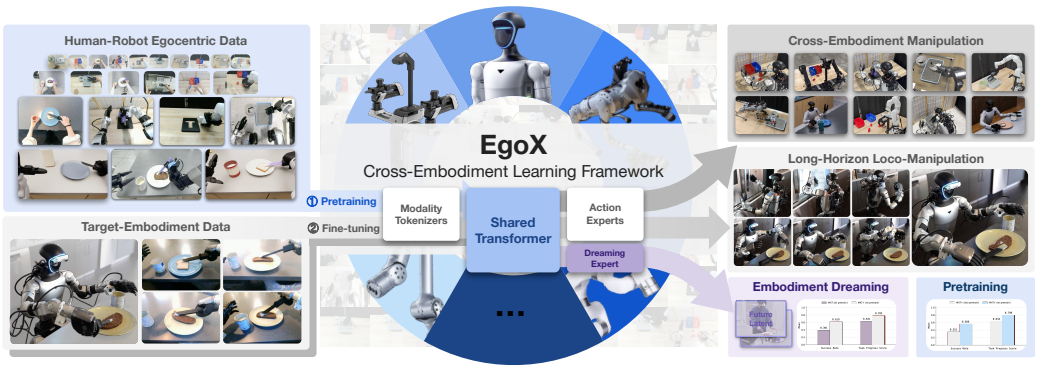

Figure 1: **Overview of EgoX.** EgoX learns from highly diverse embodiments, including humans, legged and wheeled humanoids, a quadrupedal manipulator, tabletop arms, and a simulated humanoid, with dexterous hands, grippers, and tools as end-effectors. It first pretrains on egocentric demonstrations from humans and robots (top left), then fine-tunes on demonstrations from the target embodiment (bottom left). Its model, MXT+, routes each embodiment's observations through modality tokenizers into a shared Transformer and decodes actions with embodiment-specific action experts, while a training-only dreaming expert predicts future embodiment latents (center). The resulting policies perform manipulation across embodiments and long-horizon loco-manipulation on a humanoid (right, top and middle). We find that cross-embodiment pretraining and embodiment dreaming improves task progress score over various settings (right, bottom).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Large-scale pretraining has driven rapid progress in natural language processing (Devlin et al., 2019; Brown et al., 2020) and computer vision (Radford et al., 2021; Dosovitskiy et al., 2021). Physical interaction data, however, is far less homogeneous and structured than text or images, and scaling robot manipulation requires broader embodied data than can be collected separately for every robot and task. Cross-embodiment robot learning can reuse demonstrations collected across different bodies, sensors, and action spaces (O'Neill et al., 2024; Octo Model Team et al., 2024; Kim et al., 2025; Doshi et al., 2025; Wang et al., 2024; Gemini Robotics Team et al., 2025a), while egocentric human data captures diverse physical experience in real-world settings beyond robot deployment environments (Kareer et al., 2025; Niu et al., 2025; Qiu et al., 2025; Zheng et al., 2026). Simulation offers a further source: large-scale simulated data can support policy pretraining or even zero-shot transfer (Deshpande et al., 2026; NVIDIA et al., 2025), and simulated demonstrations, including digital cousins of real tasks, can be co-trained with real-world data (Maddukuri et al., 2025; Lin et al., 2026b). However, combining these sources for low-level manipulation remains difficult because humans, robots, and simulated agents differ in morphology, end-effectors, proprioception, viewpoints, action semantics, and motion dynamics.

Recent work has examined scaling laws and data composition for robot manipulation (Lin et al., 2025; 2026a; TRI LBM Team et al., 2025), but how to scale effectively with data from diverse embodiments remains an open question. We view this as a representation and interface problem, not only a data-volume problem: humans, simulated agents, and real robots each provide distinct observations, actions, and temporal constraints. EgoX therefore treats them as embodiments on equal footing and asks how a policy can share manipulation structure while retaining the interfaces needed for each body.

We introduce EgoX, an egocentric cross-embodiment framework built around a single methodological premise: transfer should occur through semantically-corresponding modalities, while sensing and control interfaces remain embodiment-specific. The framework has three components. First, EgoX organizes human and robot demonstrations in a unified observation-action representation, expressing poses in an embodiment-centric reference frame and grouping heterogeneous signals into anatomically- or functionally-relevant modalities. Second, MXT+, a Modularized Cross-Embodiment Transformer, places embodiment-specific modality tokenizers and action experts around a shared Transformer trunk, allowing heterogeneous embodiments to update a common representation without forcing their raw observations or action spaces to match. Third, MXT+ uses *embodiment dreaming*, an auxiliary objective that predicts future embodiment latents from the current shared representation. This training-only branch encourages the trunk to encode how each embodiment's visual and proprioceptive state evolves, without requiring pixel reconstruction or deployment-time computation.

We evaluate EgoX on five long-horizon and dexterous manipulation tasks, `Kitchen-Bread`, `Kitchen-Cleanup`, `Sort`, `Insert`, and `Scoop`, across six robot embodiments: `G1-Dex5`, `xArm7-Dex4`, `G1-Sim-Dex3`, `LocoMan-Dex2`, `R1Lite-Gripper`, `Z1-Tool`. Human demonstrations provide an additional pretraining source. Our experiments compare ACT (Zhao et al., 2023), $\pi_{0.5}$ (Black et al., 2025a), and MXT+ variants across all tasks, and ablate the pretraining mixture, embodiment dreaming objective, and various architectural MXT+ design choices. The results show that cross-embodiment pretraining improves downstream task performance, that embodiment diversity matters, and that embodiment dreaming provides an additional transfer signal via the learned embodiment latent representations. In summary, our contributions are:

1. **A unified, modality-structured egocentric representation for cross-embodiment robot manipulation.** Our framework, EgoX, places human, real-robot, and simulated demonstrations in a common observation-action format while retaining embodiment-specific sensing, action-semantics, and validity masks.
2. **MXT+, A modular cross-embodiment Transformer with embodiment dreaming.** MXT+ places embodiment-specific tokenizers and action experts around a shared trunk and predicts future embodiment latents as an auxiliary training objective, to model each embodiment's transition dynamics, with no dreaming computation at deployment time.
3. **A systematic study of cross-embodiment transfer.** Experiments across five tasks and six robot embodiments show that cross-embodiment pretraining benefits downstream robot

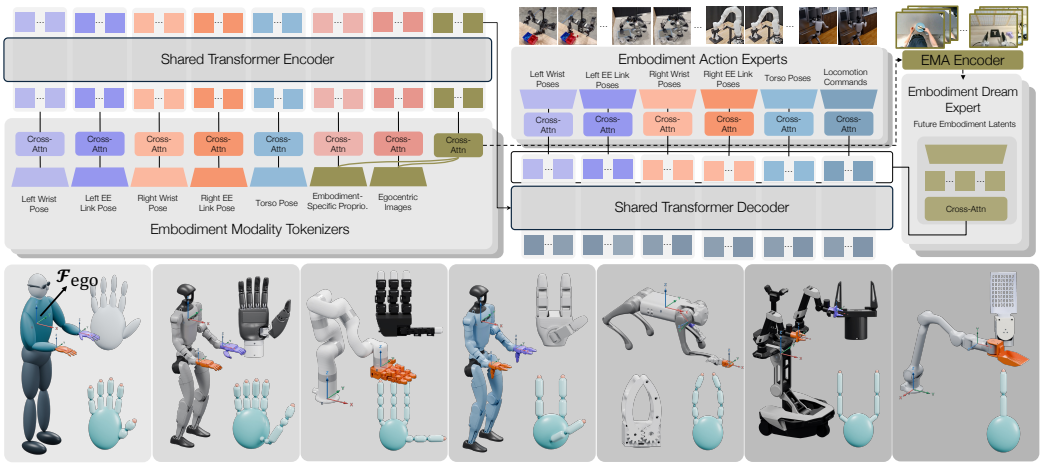

Figure 2: **MXT+ architecture and EgoX embodiments.** *Top:* embodiment-specific modality tokenizers map each embodiment's observations (egocentric images, embodiment proprioception, and wrist, end-effector-link, and torso poses) into tokens that a shared Transformer encoder-decoder trunk fuses across embodiments. Embodiment action experts decode the trunk output into embodiment-specific action chunks, while embodiment-dreaming experts predict future embodiment latents supervised by EMA copies of the corresponding modality tokenizers. *Bottom:* the embodiments used in EgoX, from left to right: `Human-Dex5`, `G1-Dex5`, `xArm7-Dex4`, `G1-Sim-Dex3`, `LocoMan-Dex2`, `R1Lite-Gripper`, and `Z1-Tool`. Each panel marks the unified egocentric frame $\mathcal{F}_{\text{ego}}$ in which wrist and end-effector poses and actions are expressed, together with its wrist frames. Left and right end-effectors are shown in purple and orange, matching the modality colors above, alongside a close-up of each end-effector and its end-effector-link keypoints.

manipulation, that broader embodiment mixtures improve transfer, and that embodiment dreaming improves the learned embodiment latent representations and provides further gains on task performance.

We discuss related work in Appendix A.1.

2 EgoX: Egocentric Cross-Embodiment Manipulation

EgoX makes three key design decisions to enable effective learning across diverse embodiments. First, we represent the heterogeneous data in a unified structure based on shared modalities within the unified egocentric frame (Sec. 2.1). To handle diverse sensing and control interfaces with different robots while facilitating positive transfer, EgoX leverages MXT+ (Fig. 2), where embodiment-specific tokenizers map observations into tokens and the transformer backbone can learn from a semantically aligned representation of diverse embodiments (Sec. 2.2). Finally, the auxiliary objective of *embodiment dreaming* encourages the transformer to learn each embodiment's dynamics (Sec. 2.3). EgoX uses a two-stage training pipeline: pretraining on mixed cross-embodiment data, and finetuning on the target embodiment task (Sec. 2.4).

2.1 Egocentric Cross-Embodiment Data Representation

Although embodiments differ substantially as whole bodies, many of their parts play corresponding roles. We exploit this by decomposing each embodiment into anatomically or functionally corresponding modalities and aligning them in a unified egocentric frame. Formally, EgoX collects demonstrations from a set of embodiments $\mathcal{E}$ that includes both humans and robots. For each embodiment $e \in \mathcal{E}$, the dataset

$$\mathcal{D}_e = \{\tau_i^e\}_{i=1}^{N_e}, \qquad \tau_i^e = \left(\boldsymbol{o}_{i,t}^e, \boldsymbol{a}_{i,t}^e\right)_{t=1}^{T_i^e}, \tag{1}$$

contains N_e trajectories. Trajectory τ_i^e has length T_i^e. The terms $\boldsymbol{o}_{i,t}^e$ and $\boldsymbol{a}_{i,t}^e$ are the observations and actions at time step t. Both are tuples of named modalities, defined below. When the trajectory is clear from context, we omit the index i.

Unified egocentric frame. To retain a consistent egocentric spatial interpretation, EgoX follows a unified egocentric frame $\mathcal{F}_{\text{ego}}$ (as shown in Fig. 2) attached to the egocentric main camera frame or the rigid body carrying the main camera, and expresses wrist and end-effector poses relative to it. At reset, the x-axis points forward toward the workspace, the y-axis points left, and the z-axis points upward. Poses streamed from teleoperation, sensing, or control interfaces are converted into this frame before data saving and model inference.

Observations and action targets. EgoX represents data from different embodiments in a unified schema of images, and proprioceptive states and actions of shared modalities. As depiced in Fig. 2, each demonstration is represented with the same observation-action dictionary structure:

$$\boldsymbol{o}_t^e := \{\boldsymbol{I}^e_{\text{main},t},\, \boldsymbol{x}^e_{\text{torso},t},\, \boldsymbol{x}^e_{\text{L/R wrist},t},\, \boldsymbol{p}^e_{\text{L/R ee link},t}, \dots\}, \tag{2}$$

$$\boldsymbol{a}_t^e := \{\boldsymbol{x}^{e,\text{target}}_{\text{torso},t},\, \boldsymbol{x}^{e,\text{target}}_{\text{L/R wrist},t},\, \boldsymbol{p}^{e,\text{target}}_{\text{L/R ee link},t}, \dots\}. \tag{3}$$

Across embodiments, observations include egocentric main-camera images $\boldsymbol{I}^e_{\text{main},t}$, torso pose proprioception $\boldsymbol{x}^e_{\text{torso},t}$, wrist pose proprioception $\boldsymbol{x}^e_{\text{L/R wrist},t}$, and end-effector link pose proprioception $\boldsymbol{p}^e_{\text{L/R ee},t}$. There may also be additional embodiment-specific streams such as wrist-camera images, robot joint states, and force feedback. Action targets are represented in the same modular format, including torso pose commands $\boldsymbol{x}^{e,\text{target}}_{\text{torso},t}$, end-effector pose commands $\boldsymbol{x}^{e,\text{target}}_{\text{L/R wrist},t}$, end-effector link pose commands $\boldsymbol{p}^{e,\text{target}}_{\text{L/R ee},t}$, and when available, velocity commands.

The end-effector link poses $\boldsymbol{p}^e_{\text{L/R ee link},t}$ are task-relevant 3D keypoints on each end-effector, expressed in its local frame, such as the fingertips of human and robot hands and the contact points of grippers and tools (bottom of Fig. 2). Because they describe end-effector geometry rather than actuator joint coordinates, hands, grippers, and tools with different kinematics are expressed in comparable geometric terms. In this unified format, human and robot demonstrations alike become egocentric trajectories that pair visual observations with torso, wrist, and end-effector-link motion.

Shared modality vocabulary. Let $\mathcal{M} = \mathcal{M}_{\text{obs}} \cup \mathcal{M}_{\text{act}}$ be the modality vocabulary shared by all embodiments, indexed by m: each entry of $\boldsymbol{o}^e_t$ and $\boldsymbol{a}^e_t$ above, with L and R counted separately, is one modality, written $\boldsymbol{o}^{e,m}_t$ or $\boldsymbol{a}^{e,m}_t$. Not every embodiment provides every modality, so $\mathcal{M}^e_{\text{obs}} \subseteq \mathcal{M}_{\text{obs}}$ and $\mathcal{M}^e_{\text{act}} \subseteq \mathcal{M}_{\text{act}}$ denote the modalities available for embodiment e. Missing modalities are handled with mask tokens and loss masking, so token sequences remain semantically aligned across embodiments (Appendix A.2).

2.2 Modularized Cross-Embodiment Transformer

To learn from heterogeneous human and robot demonstrations, EgoX uses MXT+, a Modularized Cross-Embodiment Transformer with Embodiment Dreaming. Building on the modular tokenization and action decoding of MXT (Niu et al., 2025), MXT+ separates embodiment-specific sensing and control interfaces from a shared policy backbone: observations are mapped to modality tokens, fused by a shared encoder-decoder Transformer trunk, and decoded into action chunks and future embodiment latents (Sec. 2.3) by modular experts.

Architecture overview. For each embodiment $e \in \mathcal{E}$, the MXT+ policy is the composition

$$\pi^e = \Psi^e \circ G_\theta \circ \Phi^e, \qquad \Phi^e = \{\phi^e_m\}_{m \in \mathcal{M}^e_{\text{obs}}}, \qquad \Psi^e = \{\psi^e_m\}_{m \in \mathcal{M}^e_{\text{act}}}, \tag{4}$$

where ϕ^e_m is the tokenizer of observation modality m, ψ^e_m is the action expert of action modality m, and G_θ is the shared trunk. The tokenizers Φ^e and experts Ψ^e form the paired input and output interfaces of embodiment e and are instantiated only for the modalities that e provides. Because all modalities come from the shared vocabulary $\mathcal{M}$ and missing ones are filled with mask tokens, G_θ receives the same modality layout from every embodiment. Cross-embodiment transfer therefore happens in the shared trunk, while Φ^e and Ψ^e absorb differences in sensing and control. Given an observation $\boldsymbol{o}^e_t = (\boldsymbol{o}^{e,m}_t)_{m \in \mathcal{M}^e_{\text{obs}}}$, the policy predicts an action chunk, i.e., the next H action targets of every available action modality:

$$\hat{\boldsymbol{A}}^e_t = \pi^e(\boldsymbol{o}^e_t) = \big(\hat{\boldsymbol{A}}^{e,m}_t\big)_{m \in \mathcal{M}^e_{\text{act}}}, \qquad \hat{\boldsymbol{A}}^{e,m}_t = \big(\hat{\boldsymbol{a}}^{e,m}_{t+k}\big)_{k=0}^{H-1} \in \mathbb{R}^{H \times d_m}, \tag{5}$$

where d_m is the dimension of action modality m. Appendix A.3 provides details on each module.

2.3 EMBODIMENT DREAMING

Behavioral cloning supervises the policy only through demonstrated actions; it does not require the policy to anticipate how the embodiment's own state evolves. MXT+ therefore adds *embodiment dreaming*, an auxiliary objective in which the policy predicts future *embodiment latents*, i.e., compact representations of selected future observation modalities of the active embodiment. This encourages the shared trunk to capture embodiment-specific dynamics. At the same time, supervising in latent space lets the trunk predict future embodied state at the feature level, including visual and joint configurations, without reconstructing high-dimensional observations such as images.

Embodiment latent. For embodiment e, let $\mathcal{M}^e_{\text{dream}} \subseteq \mathcal{M}^e_{\text{obs}}$ be the observation modalities used as dreaming sources: the egocentric image, robot joint configuration, and end-effector joint configuration, whichever of them e provides. Robot joint configurations are not part of the action space, but they carry embodiment-specific state that is informative about body dynamics. An embodiment tokenizer ϕ^e_{dream} adds a positional embedding to the tokens of each modality in $\mathcal{M}^e_{\text{dream}}$ and aggregates them directly into a single latent vector, without pooling over multiple output tokens. The dreaming target at a future offset k is computed by exponential-moving-average (EMA) teacher copies of the policy's modality tokenizers ϕ^e_m (Sec. 2.2) and of ϕ^e_{dream}, marked by a bar:

$$\boldsymbol{y}^e_{t+k} = \text{sg}\Big[\,\bar{\phi}^e_{\text{dream}}\Big(\big[\,\bar{\phi}^e_m\big(\boldsymbol{o}^{e,m}_{t+k}\big)\,\big]_{m\in\mathcal{M}^e_{\text{dream}}}\Big)\Big], \tag{6}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator. As in the JEPA model family (Assran et al., 2025), the stop-gradient prevents the dreaming loss from updating the target, and the EMA update in Eq. (9) makes the target evolve slowly.

Dreaming expert. A dreaming expert ψ^e_{dream} reads the same trunk output as the action experts and predicts the embodiment latents at a set of future offsets $\mathcal{K}$:

$$\big(\hat{\boldsymbol{y}}^e_{t+k}\big)_{k\in\mathcal{K}} = \big(\psi^e_{\text{dream}} \circ G_\theta \circ \Phi^e\big)\big(\boldsymbol{o}^e_t\big). \tag{7}$$

Dreaming targets are sampled from future observations by the same data pipeline as action chunks.

Latent prediction loss. The predicted latents are supervised with a cosine alignment loss and an optional magnitude loss:

$$\mathcal{L}^e_{\text{dream}} = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\Big[1 - \cos\big(\hat{\boldsymbol{y}}^e_{t+k},\, \boldsymbol{y}^e_{t+k}\big) + \beta\, ell_{\text{smooth-L1}}\big(\|\hat{\boldsymbol{y}}^e_{t+k}\|_2,\, \|\boldsymbol{y}^e_{t+k}\|_2\big)\Big], \tag{8}$$

where $\beta \geq 0$ weights the magnitude term.

EMA target update. Each teacher $\bar{\phi}$ in Eq. (6) tracks its student ϕ, i.e., ϕ^e_{dream} and ϕ^e_m for $m \in \mathcal{M}^e_{\text{dream}}$. After each optimizer step,

$$\bar{\phi} \leftarrow \alpha\,\bar{\phi} + (1-\alpha)\,\phi, \qquad \alpha \in (0,1), \tag{9}$$

where ϕ and $\bar{\phi}$ denote module parameters.

2.4 CROSS-EMBODIMENT PRETRAINING AND TARGET FINETUNING

EgoX first pretrains MXT+ on a mixture of human and robot embodiments with action prediction and embodiment dreaming, and then finetunes it on demonstrations of the target embodiment.

Training objective. For a time step t of a trajectory of embodiment e, the action loss is the chunked behavioral cloning loss over the available action modalities,

$$\mathcal{L}^e_{\text{act}} = \sum_{m\in\mathcal{M}^e_{\text{act}}} \frac{\lambda^e_m}{H}\sum_{k=0}^{H-1}\big\|\hat{\boldsymbol{a}}^{e,m}_{t+k} - \boldsymbol{a}^{e,m}_{t+k}\big\|_1, \tag{10}$$

where λ^e_m weights action modality m. Because the sum runs over $\mathcal{M}^e_{\text{act}}$, action modalities that e does not provide are excluded. In both Eq. (10) and Eq. (8), chunk steps and dreaming offsets beyond the end of the trajectory are padded and excluded from the average. The objective for embodiment e on a dataset $\mathcal{D}$ is

$$\mathcal{L}^e(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}\big[\,\mathcal{L}^e_{\text{act}} + \lambda^e_{\text{dream}}\,\mathcal{L}^e_{\text{dream}}\,\big], \tag{11}$$

where $\mathbb{E}_{\mathcal{D}}$ averages over time steps t of all trajectories in $\mathcal{D}$, and $\lambda^e_{\text{dream}} \geq 0$ weights embodiment dreaming.

Pretraining. Let $\mathcal{E}_{\text{pre}} \subseteq \mathcal{E}$ be the pretraining embodiments, including humans and multiple robots. Each batch mixes samples from one or more embodiment datasets, and MXT+ minimizes

$$\mathcal{L}_{\text{pre}} = \sum_{e \in \mathcal{E}_{\text{pre}}} \mathcal{L}^e(\mathcal{D}_e). \tag{12}$$

The shared trunk G_θ receives gradients from all embodiments, whereas the embodiment-specific modules of e are updated only by $\mathcal{D}_e$.

Finetuning. Given a target embodiment $e^\star$ and a target-task dataset $\mathcal{D}^{\text{task}}_{e^\star}$, EgoX initializes MXT+ from the pretrained checkpoint and minimizes

$$\mathcal{L}_{\text{ft}} = \mathcal{L}^{e^\star}\big(\mathcal{D}^{\text{task}}_{e^\star}\big), \tag{13}$$

where setting $\lambda^{e^\star}_{\text{dream}} = 0$ disables dreaming during finetuning. When the target task has little data, the pretrained trunk provides a manipulation prior learned from human and other-robot demonstrations, and the embodiment-specific modules adapt the shared representation to the target robot's sensors and action interface. Unlike single-source human-to-robot pretraining, EgoX does not rely on a single source or target embodiment; it co-trains on a broad embodiment mixture and then specializes to the target deployment.

Inference. At deployment, only the policy $\pi^{e^\star} = \Psi^{e^\star} \circ G_\theta \circ \Phi^{e^\star}$ is executed: it predicts action chunks $\hat{\boldsymbol{A}}^{e^\star}_t$ from egocentric observations for low-level control. The embodiment tokenizer $\phi^{e^\star}_{\text{dream}}$, the dreaming expert $\psi^{e^\star}_{\text{dream}}$, and the EMA teachers are used only during training. Algorithm 1 in Appendix A.4 summarizes the training procedure.

3 Experiments and Results

3.1 Experimental Setup

Tasks. We evaluate EgoX on five diverse manipulation tasks covering long-horizon locomotive manipulation, semantic object selection, precise alignment, tool use, and contact-rich interaction. In `Kitchen-Bread` (Sauce Spreading), the robot retrieves sauce containers from a fridge or tabletop, picks up a spreading tool, dips it into the sauce, and spreads the sauce onto bread. In `Kitchen-Cleanup` (Tool Organization), the robot picks up tools from a tabletop and places them into a container, with the tools and container initialized around a plate fixed at the center of a designated tabletop region. In `Sort` (Block Sorting), the robot sorts blocks into containers of the corresponding colors. In `Insert` (T-shaped Block Insertion), the robot picks up a T-shaped block, aligns it with its base, and inserts it. In `Scoop` (Cat Litter Scooping), the robot uses a grasped or fixed scoop to collect cat litter positioned inside a litter box, empties it into a bin, and returns the scoop to the edge of the litter box.

Embodiments. Our evaluation covers six robot embodiments (five physical robots and one simulated humanoid) that differ in morphologies, mobility, handedness, and end-effectors. `G1-Dex5`, a bimanual Unitree G1 humanoid equipped with five-finger Inspire RH56F1 dexterous hands. `xArm7-Dex4`, a tabletop xArm7 equipped with a four-finger LEAP Hand (Shaw et al., 2023); `G1-Sim-Dex3` is a simulated Unitree G1 equipped with three-finger Dex3-1 dexterous hands. `LocoMan-Dex2` (Lin et al., 2024) is a quadruped with two-finger grippers mounted on its front legs, evaluated in unimanual mode. `R1Lite-Gripper`, a Galaxea R1 Lite wheeled bimanual mobile manipulator with parallel-jaw grippers. `Z1-Tool`, a tabletop Unitree Z1 arm with a fixed, non-actuated scoop. Egocentric human demonstrations (`Human-Dex5 Pro`) provide an additional embodiment for cross-embodiment pretraining. Together, these platforms span dexterous hands, grippers, and fixed tools across diverse robot morphologies and mobility configurations.

Baselines and training settings. We compare MXT+ with ACT (Zhao et al., 2023) and $\pi_{0.5}$ (Black et al., 2025a), and evaluate MXT+ both with and without cross-embodiment pretraining. ACT is trained directly on demonstrations from the target embodiment and task. *MXT+ without pretraining* is likewise trained from scratch on target-embodiment demonstrations, using the same modular

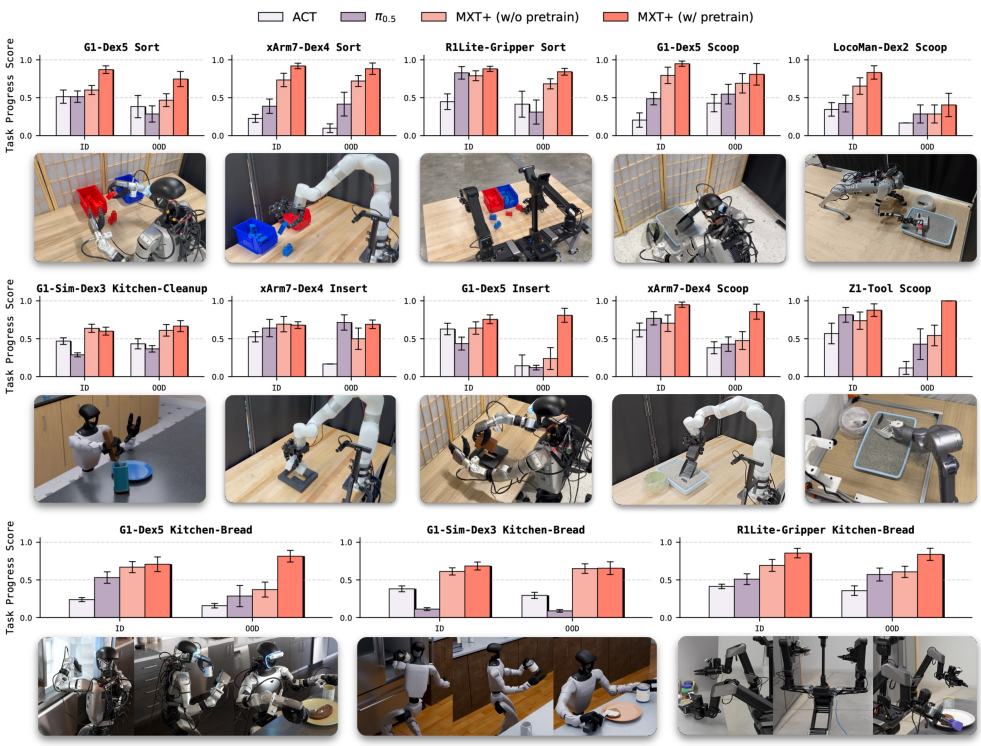

Figure 3: **Main results.** Normalized task progress score of ACT, $\pi_{0.5}$, MXT+ without pretraining, and MXT+ with cross-embodiment pretraining on 13 task–embodiment pairs spanning five tasks (`Kitchen-Bread`, `Kitchen-Cleanup`, `Sort`, `Insert`, `Scoop`) and six embodiments (five physical robots and the simulated `G1-Sim-Dex3`). Each panel reports ID and OOD performance.

tokenizers, shared Transformer trunk, and action experts as the pretrained variant. *MXT+ with pretraining* is first trained on a mixture of egocentric demonstrations from humans, physical robots, and simulated robots, and then fine-tuned on the target embodiment and task following Section 2.4. The comparison between these two MXT+ variants evaluates the benefit of cross-embodiment pretraining without changing the downstream policy architecture, while ACT and $\pi_{0.5}$ serve as baselines representing Transformer-based imitation learning and pretrained VLA models finetuned on target-embodiment demonstrations, respectively. We additionally study the sources of transfer through pretraining-mixture and model ablations. Specifically, we vary the embodiments included in pretraining to examine the effects of embodiment diversity and composition, compare models with and without embodiment dreaming and assess the contribution of the modular design.

Evaluation settings and metrics. For each task and embodiment, we assess policies under in-distribution (ID) and out-of-distribution (OOD) conditions, defined relative to the target-embodiment training demonstrations. ID settings use objects, initial configurations, and environmental conditions similar to those encountered during target training, while OOD settings introduce task-specific held-out variations, such as novel object instances, appearances, poses, or scene configurations. Each real-world task–embodiment pair is evaluated over 20 trials (13 ID and 7 OOD), and each simulated pair over 50 trials (32 ID and 18 OOD). We report the *average normalized task progress score*, which credits partial completion of intermediate task stages and normalizes the accumulated progress by the task's maximum score.

3.2 Experimental Evaluation

Our evaluation aims to address the following questions: (1) Does *egocentric cross-embodiment pretraining* improve downstream dexterous manipulation compared to training from scratch on target-embodiment data? (2) How do the *diversity and composition* of pretraining embodiments affect real-robot performance? (3) Does *embodiment dreaming* improve cross-embodiment learning? (4)

Does Latent Geometry Become More Similar Across Embodiments? (5) Does *modular design* enhance the cross-embodiment transfer?

Does egocentric cross-embodiment pretraining outperform training from scratch? We examine whether egocentric cross-embodiment pretraining improves downstream policy learning compared with training the same policy architecture from scratch on target-embodiment demonstrations. In Fig. 3, cross-embodiment pretraining consistently improves the task progress score of MXT+.

Under ID evaluation, cross-embodiment pretraining yields particularly large gains on `G1-Dex5 Sort`, `LocoMan-Dex2 Scoop`, and `xArm7-Dex4 Scoop`. The benefit becomes even more pronounced under OOD evaluation, with substantial improvements on `G1-Dex5 Kitchen-Bread`, `G1-Dex5 Insert`, and `Z1-Tool Scoop`, where pretrained MXT+ substantially closes the gap to or exceeds the performance of other baselines.

Overall, these results suggest that cross-embodiment pretraining on mixed human-robot demonstrations exposes the policy to a broader range of visual, geometric, and interaction patterns than target-task demonstrations alone, contributing to improved downstream generalization.

Does pretraining benefit from embodiment diversity? We next study how the composition of the pretraining mixture affects downstream performance. As shown in Fig. 4, we compare four settings: no pretraining, human-only pretraining, human plus other-robot pretraining, and full cross-embodiment pretraining. Human-only pretraining uses egocentric human demonstrations; human plus other-robot pretraining additionally incorporates demonstrations from robot embodiments other than the target robot; and full cross-embodiment pretraining further includes demonstrations from the target embodiment on non-target tasks. All variants are fine-tuned on the same target-embodiment, target-task demonstrations. We evaluate this ablation across both real-robot and simulation settings, covering `Sort` and `Insert` on `R1 Lite-Gripper`, `G1-Dex5` and `xArm7-Dex4`, as well as `Kitchen-Bread` and `Kitchen-Cleanup` on simulated `G1-Sim-Dex3`.

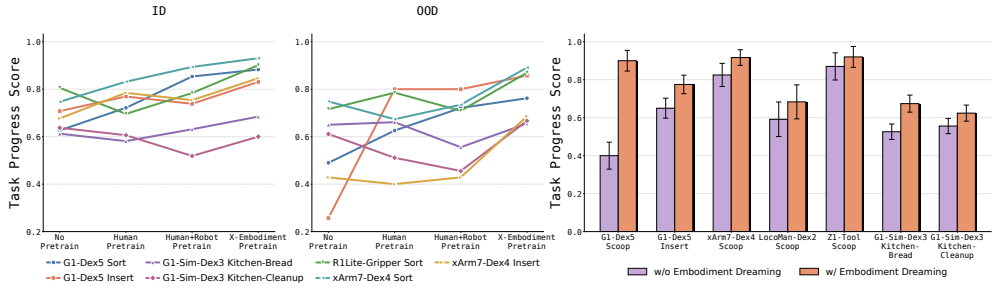

Figure 4: **Pretraining data composition and embodiment dreaming ablations.** *Left* and *middle*: task progress scores under ID and OOD evaluation as we vary the pretraining mixture: no pretraining, human-only, human plus robot, and full cross-embodiment. *Right:* task progress scores with and without embodiment dreaming across tasks and embodiments.

Human-only pretraining can improve performance over no pretraining on several task–embodiment pairs, suggesting that egocentric human demonstrations can provide large-scale transferable manipulation priors despite the embodiment gap. Adding demonstrations from other robot embodiments can provide further gains in some settings, although the effect varies across tasks and evaluation conditions. Overall, incorporating a broader range of embodiments into the pretraining mixture tends to improve performance, with full cross-embodiment pretraining showing the strongest overall trend, particularly under OOD evaluation. More broadly, cross-embodiment transfer can become more consistent and yield stronger gains as the diversity of tasks and embodiments increases, highlighting the value of broad and diverse pretraining data.

Does *embodiment dreaming* improve cross-embodiment learning? Finally, we study the effect of embodiment dreaming by comparing MXT+ with MXT, its counterpart without the embodiment-dreaming objective. We evaluate this ablation across five embodiments and seven task–embodiment settings, including `Scoop` on four embodiments (`G1-Dex5`, `xArm7-Dex4`, `LocoMan-Dex2`, and `Z1-Tool`), `Insert` on `G1-Dex5`, and both `Kitchen-Bread` and `Kitchen-Cleanup` on `G1-Sim-Dex3`. The gains are observed across different tasks and embodiments, with particu-

larly pronounced improvements on `G1-Dex5`, suggesting that embodiment dreaming provides an additional learning signal beyond action prediction, improving downstream manipulation performance. These results provide evidence that embodiment dreaming canimprove representations for cross-embodiment learning, facilitating knowledge transfer across substantially different kinematics, viewpoints, and action spaces.

Does Latent Geometry Become More Similar Across Embodiments? We compute Gromov–Wasserstein (GW) distances between embodiment-specific distributions of shared-trunk latent representations extracted from `Scoop` demonstrations.

Each embodiment's demonstrations are encoded by its corresponding checkpoint, with matched samples across conditions. GW uses median-normalized within-distribution Euclidean distances in the original latent space. We report percentage changes in GW relative to each baseline, with negative values indicating greater geometric similarity. We compare MXT+ with and without cross-embodiment pretraining. To assess embodiment dreaming, we compare MXT with MXT+, both pretrained and finetuned. Fig. 5 shows that pretraining lowers GW for 9 of 10 embodiment pairs and reduces mean GW across the pairs by 21.30%. MXT+ yields lower GW than MXT for all 10 pairs, reducing mean GW across the pairs by 38.52%. These findings suggest more similar internal distance structures among embodiment-specific latent representations for the same manipulation task.

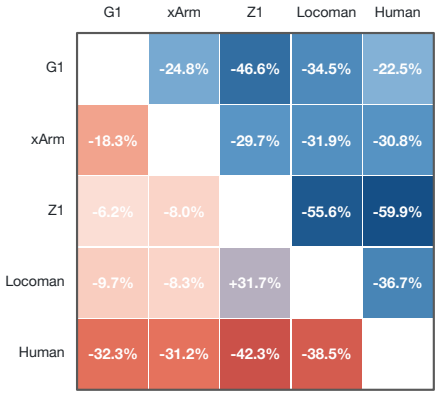

Figure 5: GW change across embodiment pairs. **Pretraining**: red, lower triangle; **Embodiment dreaming**: blue, upper triangle.

Does *modular design* enhance the cross-embodiment transfer? We next examine the role of our modular architecture by comparing MXT+ with a coarse variant that merges modality-specific inputs and actions. Fig. 6 evaluates both variants with and without cross-embodiment pretraining on `Z1-Tool Scoop`, `G1-Sim-Dex3 Kitchen-Bread`, and `G1-Sim-Dex3 Kitchen-Cleanup`. The modular design consistently outperforms the coarse variant, with particularly pronounced gains for the higher-dimensional G1-Sim-Dex3, indicating that modality-specific interfaces benefit both downstream learning and cross-embodiment transfer.

The benefit of modularity is especially clear on `Kitchen-Cleanup`: after pretraining, the coarse variant achieves zero task progress, while modular MXT+ retains substantial performance. The first step of this task requires a 3D velocity command for locomotion, a small but critical modality within G1's 51-dimensional action space. Treating all actions as a single coarse representation can hinder learning this modality, whereas the modular design provides dedicated experts for each action modality. Overall, these results show that modularity is crucial for both downstream performance and effective cross-embodiment transfer, particularly for high-dimensional embodiments with diverse modalities.

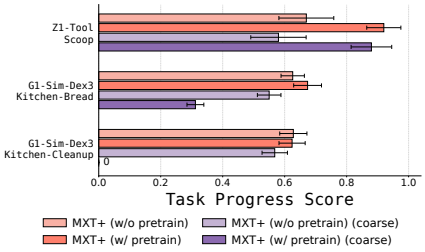

Figure 6: **Coarse ablation.** MXT+ task progress scores with/without the coarse variant and pretraining.

4 Conclusion

We presented EgoX, an egocentric cross-embodiment framework that represents human, real-robot, and simulated demonstrations in a unified, modality-structured format and trains MXT+, a Modularized Cross-Embodiment Transformer with Embodiment Dreaming. Embodiment-specific tokenizers and action experts preserve each embodiment's sensing and control interfaces around a shared trunk, and embodiment dreaming trains this trunk to predict future embodiment latents, modeling each body's dynamics at no inference cost. Across five tasks and six embodiments, cross-embodiment pretraining improves downstream manipulation, broader embodiment mixtures improve transfer, and embodiment dreaming yields further gains, especially on contact-rich tool use.

AI USE STATEMENT

We used AI coding agents to implement parts of our codebase. The authors verified the resulting code and tested it on physical robots. We also used these agents to generate 3D assets for simulation, assist with visual domain randomization for the simulation dataset, process and analyze data from physical and simulated experiments, tune hyperparameters and monitor training of the model, and render 3D robot models for figures. We did not use AI tools to generate the research ideas or design the methodology of this work. The authors take responsibility for the final content of this work.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-RDT: Human manipulation enhanced bimanual robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18135–18143, 2026. doi: 10.1609/aaai.v40i22.38875. URL https://ojs.aaai.org/index.php/AAAI/article/view/38875.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems (RSS)*, 2025b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Robotics: Science and Systems (RSS)*, 2025. doi: 10.15607/RSS.2025.XXI.014. URL https://www.roboticsproceedings.org/rss21/p014.html. arXiv:2505.06111.

Lawrence Yunliang Chen, Karthik Dharmarajan, Kush Hari, Chenfeng Xu, Quan Vuong, and Ken Goldberg. MIRAGE: Cross-embodiment zero-shot policy transfer with cross-painting. In *Robotics: Science and Systems (RSS)*, 2024.

Sirui Chen, Chen Wang, Kaden Nguyen, Li Fei-Fei, and C Karen Liu. ARCap: Collecting High-quality Human Demonstrations for Robot Learning with Augmented Reality Feedback. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8291–8298. IEEE, 2025.

Abhay Deshpande, Maya Guru, Rose Hendrix, Snehal Jauhri, Ainaz Eftekhar, Rohun Tripathi, Max Argus, Jordi Salvador, Haoquan Fang, Matthew Wallingford, et al. MolmoB0T: Large-scale simulation enables zero-shot manipulation. *arXiv preprint arXiv:2603.16861*, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning*

Research, pp. 496–512. PMLR, 2025. URL https://proceedings.mlr.press/v270/doshi25a.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yankai Fu, Ning Chen, Junkai Zhao, Shaozhe Shan, Guocai Yao, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. METIS: Multi-Source Egocentric Training for Integrated Dexterous Vision-Language-Action Model. *arXiv preprint arXiv:2511.17366*, 2025. doi: 10.48550/arXiv.2511.17366. URL http://arxiv.org/abs/2511.17366.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. DreamDojo: A real-time robot world model from large-scale human videos. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, et al. Gemini Robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025a.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025b.

Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World Models for Learning Dexterous Hand-Object Interactions from Human Videos. *arXiv preprint arXiv:2512.13644*, 2025. doi: 10.48550/arXiv.2512.13644. URL http://arxiv.org/abs/2512.13644.

Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Marion Lepert, Ria Doshi, and Jeannette Bohg. SHADOW: Leveraging segmentation masks for cross-embodiment policy transfer. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3536–3550. PMLR, 2025. URL https://proceedings.mlr.press/v270/lepert25a.html.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World Action Models Beyond Pixels with In-the-Wild Egocentric Human Data. *arXiv preprint arXiv:2607.08436*, 2026a. doi: 10.48550/arXiv.2607.08436. URL https://arxiv.org/abs/2607.08436.

Boyu Li, Chaoyi Xu, Haoqi Yuan, Xinrun Xu, Börje F. Karlsson, Dongbin Zhao, Haoran Li, and Zongqing Lu. X-DiffVLA: X-embodied diffusion action heads for vision-language-action models. In *Robotics: Science and Systems (RSS)*, 2026b.

Changyi Lin, Xingyu Liu, Yuxiang Yang, Yaru Niu, Wenhao Yu, Tingnan Zhang, Jie Tan, Byron Boots, and Ding Zhao. LocoMan: Advancing versatile quadrupedal dexterity with lightweight loco-manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6877–6884. IEEE, 2024.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *International Conference on Learning Representations (ICLR)*, 2025.

Fanqi Lin, Kushal Arora, Jean Mercat, Haruki Nishimura, Paarth Shah, Chen Xu, Mengchao Zhang, Mark Zolotas, Maya Angeles, Owen Pfannenstiehl, Andrew Beaulieu, and Jose Barreiros. A systematic study of data modalities and strategies for co-training large behavior models for robot manipulation. In *Robotics: Science and Systems (RSS)*, 2026a.

Kevin Lin, Ajay Mandlekar, Caelan Reed Garrett, Nikita Chernyadev, Yu Fang, Runyu Ding, Yuqi Xie, Justin Tran, Linxi Fan, and Yuke Zhu. HumanoidMimicGen: Data generation for loco-manipulation via whole-body planning. *arXiv preprint arXiv:2605.27724*, 2026b.

Yunfan Lou, Yifan Ye, Yankai Fu, Jun Cen, Xiaowei Chi, Yaoxu Lyu, Peidong Jia, Sirui Han, Zhihe Lu, and Shanghang Zhang. Dream-Tac: A unified tactile world action model for contact-rich robot manipulation. *arXiv preprint arXiv:2606.08737*, 2026.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-H0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026a. URL https://arxiv.org/abs/2601.12993.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-H0.7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026b.

Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, Scott Reed, Ken Goldberg, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. In *Robotics: Science and Systems (RSS)*, 2025.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205 of *Proceedings of Machine Learning Research*, pp. 892–909. PMLR, 2023. URL https://proceedings.mlr.press/v205/nair23a.html.

Yaru Niu, Yunzhe Zhang, Mingyang Yu, Changyi Lin, Chenhao Li, Yikai Wang, Yuxiang Yang, Wenhao Yu, Tingnan Zhang, Zhenzhen Li, Jonathan Francis, Bingqing Chen, Jie Tan, and Ding Zhao. Human2LocoMan: Learning versatile quadrupedal manipulation with human pretraining. In *Robotics: Science and Systems (RSS)*, 2025.

Yaru Niu, Zhenlong Fang, Binghong Chen, Shuai Zhou, Revanth Krishna Senthilkumaran, Hao Zhang, Bingqing Chen, Chen Qiu, H Eric Tseng, Jonathan Francis, and Ding Zhao. Learning versatile humanoid manipulation with touch dreaming. *arXiv preprint arXiv:2604.13015*, 2026.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Austin Patel and Shuran Song. GET-Zero: Graph embodiment transformer for zero-shot embodiment generalization. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14262–14269, 2025.

Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y. Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. EgoBridge: Domain adaptation for generalizable imitation from egocentric human data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 47092–47122, 2025. doi: 10.52202/085713-1569. URL https://papers.neurips.cc/paper_files/paper/2025/hash/434756a76c44e220d5f32d6e0b488c0c-Abstract-Conference.html.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. EgoVerse: An egocentric human dataset for robot learning from around the world. In *Robotics: Science and Systems (RSS)*, 2026.

Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, et al. Humanoid policy $\sim$ human policy. In *Proceedings of The 9th Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2888–2906. PMLR, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Carmelo Sferrazza, Dun-Ming Huang, Fangchen Liu, Jongmin Lee, and Pieter Abbeel. Body transformer: Leveraging robot embodiment for policy learning. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3407–3424. PMLR, 2025. URL https://proceedings.mlr.press/v270/sferrazza25a.html.

Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. LEAP Hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In *Robotics: Science and Systems (RSS)*, 2023.

Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Tianyu Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. In *Robotics: Science and Systems (RSS)*, 2026. doi: 10.15607/RSS.2026.XXII.204. URL https://www.roboticsproceedings.org/rss22/p204.html.

Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. HRP: Human affordances for robotic pre-training. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

TRI LBM Team, Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, et al. A careful examination of large behavior models for multitask dexterous manipulation. *arXiv preprint arXiv:2507.05331*, 2025.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 124420–124450, 2024.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, Sheng Zang, Weiduo Yuan, Marco Pavone, Di Huang, and Yue Wang. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. In *Robotics: Science and Systems (RSS)*, 2026a. doi: 10.15607/RSS.2026.XXII.021. URL https://www.roboticsproceedings.org/rss22/p021.html.

Zhenyu Wei, Yunchao Yao, and Mingyu Ding. One hand to rule them all: Canonical representations for unified dexterous manipulation. In *Robotics: Science and Systems (RSS)*, 2026b.

Xiaomeng Xu, Jisang Park, Han Zhang, Eric Cousineau, Aditya Bhat, Jose Barreiros, Dian Wang, Jeannette Bohg, and Shuran Song. HoMMI: Learning whole-body mobile manipulation from human demonstrations. In *Robotics: Science and Systems (RSS)*, 2026.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. EgoVLA: Learning Vision-Language-Action Models from Egocentric Human Videos. *arXiv preprint arXiv:2507.12440*, 2025. URL https://arxiv.org/abs/2507.12440.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Chengbo Yuan, Rui Zhou, Mengzhen Liu, Yingdong Hu, Shengjie Wang, Li Yi, Chuan Wen, Shanghang Zhang, and Yang Gao. MotionTrans: Human VR Data Enable Motion-Level Learning for Robotic Manipulation Policies. *arXiv preprint arXiv:2509.17759*, 2025. URL https://arxiv.org/abs/2509.17759.

Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. CLAP: Contrastive Latent Action Pretraining for Learning Vision-Language-Action Models from Human Videos. *arXiv preprint arXiv:2601.04061*, 2026. doi: 10.48550/arXiv.2601.04061. URL http://arxiv.org/abs/2601.04061.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, et al. FLARE: Robot learning with implicit world modeling. In *Proceedings of The 9th Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3952–3971. PMLR, 2025. URL https://proceedings.mlr.press/v305/zheng25a.html.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

Lawrence Y. Zhu, Pranav Kuppili, Ryan Punamiya, Patcharapong Aphiwetsa, Dhruv Patel, Simar Kareer, Sehoon Ha, and Danfei Xu. EMMA: Scaling Mobile Manipulation via Egocentric Human Data. *IEEE Robotics and Automation Letters*, 11(3):3087–3094, 2026.

A Appendix

A.1 Related Work

Cross-Embodiment Learning for Robotic Manipulation. Cross-embodiment robot learning aims to train policies that benefit from data collected on different bodies, sensors, and control interfaces. Large robot datasets and generalist policies demonstrate the value of scaling policy learning across tasks and robot embodiments (O'Neill et al., 2024; Kim et al., 2025; Black et al., 2025b;a; Gemini Robotics Team et al., 2025b), and recent work extends such pretraining to include human data (Luo et al., 2026a; NVIDIA et al., 2025). Prior methods handle embodiment heterogeneity with shared-trunk policies that use embodiment-specific tokenizers or action heads (Doshi et al., 2025; Niu et al., 2025; Wang et al., 2024), with a single embodiment-conditioned action head (Li et al., 2026b), or with image editing that aligns robot appearance across embodiments (Chen et al., 2024; Lepert et al., 2025). Existing methods either cover a relatively narrow range of embodiments and end-effectors or preserve heterogeneity by giving each embodiment a separate interface, leaving the extent of modality-level sharing unclear. In contrast, EgoX covers highly heterogeneous embodiments and end-effectors: it is evaluated on humanoid, wheeled bimanual, tabletop-arm, quadrupedal, and simulated platforms, with three- to five-fingered dexterous hands, two-finger and parallel-jaw grippers, and a tool. MXT+ combines embodiment-specific interfaces with modality-level sharing through modular tokenizers and action experts, and adds embodiment dreaming as a training signal through the prediction of future embodiment latents.

Robot Learning from Egocentric Data. Egocentric human data offers a scalable source of manipulation experience spanning diverse tasks, objects, and environments. Recent advances in wearable devices, such as XR headsets and smart glasses, have improved the quality and scale of such data (Chen et al., 2025; Hoque et al., 2026; Punamiya et al., 2026). Human videos have been used to learn transferable priors through visual, affordance, and latent-action objectives (Nair et al., 2023; Srirama et al., 2024; Ye et al., 2025; Bu et al., 2025; Zhang et al., 2026). Other methods use human demonstrations directly for policy learning, through domain adaptation (Punamiya et al., 2025), pre-training or co-training (Kareer et al., 2025; Yuan et al., 2025; Yang et al., 2025; Zheng et al., 2026; Shi et al., 2026; Zhu et al., 2026; Bi et al., 2026; Wei et al., 2026a), and transfer without additional target-robot demonstrations (Xu et al., 2026). Much of this literature emphasizes transferring human experience to target robot embodiments, while EgoX focuses on collecting and jointly learning from egocentric demonstrations across human and heterogeneous robot embodiments, and investigates how this shared training benefits manipulation across diverse platforms.

World Action Models and Embodiment Modeling. World models learn to predict how the world evolves under actions (Assran et al., 2025; Goswami et al., 2025; Gao et al., 2026). World action models couple such prediction with policy learning by jointly predicting actions and future observations, with targets including pixels, pretrained visual features, and 3D motion flow (Li et al., 2026a), embeddings from frozen or separately pretrained encoders (Zheng et al., 2025; Luo et al., 2026b), and future tactile signals or tactile latents (Niu et al., 2026; Lou et al., 2026). These works show that predicting future observations provides a useful training signal for policy learning. Other methods model the embodiment itself, either by predicting its dynamics or kinematics (Fu et al., 2025; Patel & Song, 2025), or by encoding its structure through body graphs, canonical morphology representations, or embodiment-specific functional components in action generation (Sferrazza et al., 2025; Wei et al., 2026b; Li et al., 2026b). EgoX combines the ingredients of world action models and embodiment modeling through embodiment dreaming: alongside actions, MXT+ predicts future embodiment latents from egocentric visual features, robot joint configurations, and end-effector joint configurations using exponential-moving-average copies of its own modality encoders. The objective helps the shared trunk learn embodiment-aware dynamics while leaving the deployment-time action policy unchanged.

A.2 Handling Missing Modalities

Not every embodiment provides every modality in the shared vocabulary $\mathcal{M}$ (Sec. 2.1); for example, single-arm embodiments have no left wrist or left end-effector link. EgoX handles such gaps so that token sequences remain semantically aligned across embodiments. For each embodiment e, the modality indicator

$$\delta_m^e = \mathbb{I}\,[\,m \text{ is available for } e\,] \in \{0, 1\} \tag{14}$$

defines the available observation modalities $\mathcal{M}^e_{\text{obs}} = \{m \in \mathcal{M}_{\text{obs}} : \delta^e_m = 1\}$, and $\mathcal{M}^e_{\text{act}}$ is defined analogously. Unavailable observation modalities are replaced by learnable mask tokens (Appendix A.3), so the shared trunk always receives the same modality layout, and unavailable action modalities are excluded from the action loss (Sec. 2.4).

A.3 Details on MXT+

Modality tokenizers. Each tokenizer ϕ^e_m maps modality m to a fixed number L_m of d-dimensional tokens. Modalities are defined at an anatomical or functional level, e.g., egocentric vision, body pose, end-effector pose, end-effector joint pose, and end-effector keypoints, so embodiments with different morphologies expose compatible interfaces whenever they share a modality. Image modalities are encoded by a visual backbone f_{vis} shared across all embodiments, followed by an embodiment-specific cross-attention aggregator g^e_m, so that $\phi^e_m = g^e_m \circ f_{\text{vis}}$. State modalities are encoded by lightweight MLPs followed by cross-attention aggregation. Unavailable modalities are replaced by mask tokens (Appendix A.2), and the tokens of all modalities are concatenated in a fixed modality order:

$$\boldsymbol{Z}^e_t = \left[\, \boldsymbol{Z}^{e,m}_t \,\right]_{m \in \mathcal{M}_{\text{obs}}} \in \mathbb{R}^{L \times d}, \qquad L = \textstyle\sum_{m \in \mathcal{M}_{\text{obs}}} L_m. \tag{15}$$

Tokenizing per modality preserves each modality's semantic identity and avoids mixing heterogeneous signals too early. The trunk therefore always consumes the same token interface, while the tokenizers absorb embodiment-specific differences in dimensionality, camera availability, and kinematic parameterization.

Shared Transformer trunk. The trunk G_θ is an encoder-decoder Transformer. The encoder applies self-attention over $\boldsymbol{Z}^e_t$, and the decoder cross-attends from N_q learnable query tokens $\boldsymbol{Q} \in \mathbb{R}^{N_q \times d}$ to the encoder outputs:

$$\boldsymbol{U}^e_t = G_\theta\big(\boldsymbol{Z}^e_t; \boldsymbol{Q}\big) \in \mathbb{R}^{N_q \times d}. \tag{16}$$

Because every embodiment is presented through the same token interface, G_θ is trained on all embodiments without requiring their raw observations or actions to share a dimensionality. The trunk is therefore the main conduit for cross-embodiment transfer, while the tokenizers and experts retain the embodiment-specific details needed for execution.

Modular action experts. For each available action modality $m \in \mathcal{M}^e_{\text{act}}$, the action expert decodes the corresponding action chunk from the trunk output:

$$\hat{\boldsymbol{A}}^{e,m}_t = \psi^e_m\big(\boldsymbol{U}^e_t\big) \in \mathbb{R}^{H \times d_m}. \tag{17}$$

Action modalities are drawn from the shared vocabulary $\mathcal{M}_{\text{act}}$, so anatomically or functionally corresponding outputs occupy the same modality across embodiments. For example, the end-effector pose of a single-arm robot corresponds to the right end-effector pose of a bimanual embodiment. The experts remain embodiment-specific to handle differences in dimensionality, control semantics, and validity masks, while the structure shared across embodiments is learned by the trunk.

A.4 Training Algorithm

Algorithm 1 details the full EgoX training procedure, including cross-embodiment pretraining, target finetuning, embodiment dreaming, and the EMA target update.

A.5 Embodiment and Hardware Details

EgoX is evaluated on five physical robot embodiments spanning substantially different morphologies, mobility configurations, and manipulation interfaces, as shown in Fig. 7. The platforms include a bimanual humanoid with dexterous hands (G1-DEX5), a wheeled bimanual mobile manipulator with parallel-jaw grippers (R1LITE-GRIPPER), a tabletop manipulator with a dexterous hand (XARM7-DEX4), a fixed-base arm equipped with a non-actuated tool (Z1-TOOL), and a quadrupedal manipulator with a two-finger gripper (LOCOMAN-DEX2).

Across all embodiments, EgoX represents manipulation actions using a shared set of semantically corresponding geometric action modalities. These include end-effector pose targets and, when available, task-relevant end-effector keypoint targets. For dexterous hands and grippers, the keypoints

Algorithm 1 EgoX pretraining and finetuning with embodiment dreaming

Input: Datasets $\{\mathcal{D}_e\}_{e\in\mathcal{E}_{\text{pre}}}$, target dataset $\mathcal{D}^{\text{task}}_{e^\star}$
Input: Action horizon H, dreaming offsets $\mathcal{K}$, EMA decay α
Output: Target policy $\pi^{e^\star} = \Psi^{e^\star} \circ G_\theta \circ \Phi^{e^\star}$

Initialize tokenizers, trunk G_θ, action experts, and dreaming experts
Initialize each teacher $\bar{\phi}$ as a copy of its student ϕ
for each pretraining step **do**
  Sample a batch from $\{\mathcal{D}_e\}_{e\in\mathcal{E}_{\text{pre}}}$
  Predict action chunks $\hat{\boldsymbol{A}}^e_t$ and embodiment latents $\hat{\boldsymbol{y}}^e_{t+k}$
  Compute dreaming targets $\boldsymbol{y}^e_{t+k}$ with Eq. (6)
  Compute $\mathcal{L}_{\text{pre}}$ with Eq. (12) and update students by gradient descent
  Update teachers with Eq. (9)
end for
Initialize MXT+ from the pretrained checkpoint
for each finetuning step **do**
  Sample a batch from $\mathcal{D}^{\text{task}}_{e^\star}$
  Compute $\mathcal{L}_{\text{ft}}$ with Eq. (13) and update students by gradient descent
  Update teachers with Eq. (9) if $\lambda^{e^\star}_{\text{dream}} > 0$
end for
return $\pi^{e^\star}$

correspond to fingertips, while embodiments that do not provide such structure can use only the applicable subset of action modalities. The predicted geometric targets are mapped to each embodiment's native control space through embodiment-specific retargeting and low-level control interfaces.

G1-Dex5. G1-DEX5 is a Unitree G1 humanoid equipped with two Inspire RH56F1 five-finger dexterous hands. EgoX can predict body pose, left/right end-effector pose targets together with five fingertip targets for each hand and optional velocity command for whole robot. These geometric targets of hand and e are retargeted to the corresponding arm and hand control spaces for execution. For locomotion and whole-body stabilization, we use the low-level whole-body controller from Humanoid Touch Dream (Niu et al., 2026), which is kept fixed throughout our experiments.

R1Lite-Gripper. R1LITE-GRIPPER is a Galaxea R1 Lite wheeled bimanual mobile manipulator equipped with parallel-jaw grippers. EgoX predicts left/right end-effector pose targets together with two fingertip targets for each gripper and optional velocity command. These geometric targets are

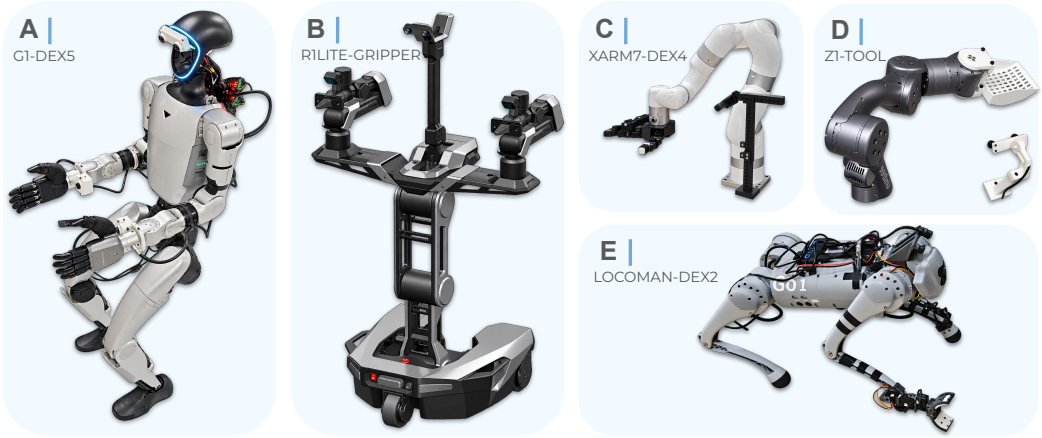

Figure 7: Physical robot embodiments used in EgoX.

Table 1: **Embodiment-specific interfaces used in EgoX.** EgoX represents manipulation actions using semantically corresponding geometric modalities while allowing each embodiment to expose only the modalities applicable to its morphology. EEF denotes end effector.

Embodiment	Side	Robot Joint Dim.	EEF Joint Dim.	Body Action	EEF Pose Action	Fingertip Targets
G1-DEX5	Bimanual	29	$6+6$	Yes	L/R	5 / hand
R1LITE-GRIPPER	Bimanual	15	$1+1$	No	L/R	2 / gripper
XARM7-DEX4	Right	7	16	No	Right	4
LOCOMAN-DEX2	Right	18	1	Yes	Right	2
Z1-TOOL	Right	6	–	No	Right	–

"Robot Joint Dim." and "EEF Joint Dim." denote the native proprioceptive dimensions used by each embodiment. EgoX predicts geometric EEF-pose and, when applicable, fingertip targets, which are subsequently mapped to each embodiment's native control space.

mapped to the native arm and gripper control spaces of the R1 Lite for execution. Mobile-base commands are handled through the embodiment-specific locomotion interface.

xArm7-Dex4. XARM7-DEX4 consists of an xArm7 tabletop manipulator equipped with a four-finger LEAP Hand (Shaw et al., 2023). EgoX predicts an end-effector pose target together with four fingertip targets. These geometric targets are retargeted to the xArm7 arm and LEAP Hand control spaces for execution. Since the platform has a fixed base, it does not use a locomotion action modality.

Z1-Tool. Z1-TOOL is a Unitree Z1 tabletop manipulator equipped with a fixed, non-actuated scoop. For this embodiment, EgoX predicts only the end-effector pose target, which is executed by the underlying arm controller. Since the scoop is rigidly attached to the arm and has no actuated joints, this embodiment does not use a separate end-effector keypoint or actuation modality.

LocoMan-Dex2. LOCOMAN-DEX2 is a quadrupedal manipulation platform with two-finger grippers mounted on its front legs (Lin et al., 2024). We evaluate the platform in a unimanual configuration. EgoX predicts an end-effector pose target together with two fingertip targets for the active gripper. These geometric targets are mapped to the native control space of the active manipulation limb for execution.

A.6 DATASET AND DATA COLLECTION

Dataset overview. EgoX is trained on a heterogeneous collection of egocentric manipulation demonstrations spanning human, physical-robot, and simulated embodiments. In total, the dataset contains 4,851 trajectories corresponding to 34.38 hours of interaction. Table 2 summarizes the dataset composition for each embodiment, while Fig. 8 visualizes the task distribution across embodiments.

Table 2: **Dataset statistics.** Number of training and validation trajectories and total interaction duration for each embodiment used in EgoX.

Embodiment	Domain	Tasks	Train	Val.	Total	Hours
HUMAN-DEX5	Human	Insert, Kitchen, Scoop, Sort	2,198	198	2,396	9.26
G1-DEX5	Real	Insert, Kitchen, Scoop, Sort	622	88	710	6.70
XARM7-DEX4	Real	Insert, Scoop, Sort	398	39	437	3.64
G1-SIM-DEX3	Sim	Kitchen, Kitchen-Spreader	691	58	749	7.51
LOCOMAN-DEX2	Real	Scoop	34	3	37	0.35
R1LITE-GRIPPER	Real	Insert, Kitchen, Sort	398	40	438	6.36
Z1-TOOL	Real	Scoop	77	7	84	0.56
Total	–	–	**4,418**	**433**	**4,851**	**34.38**

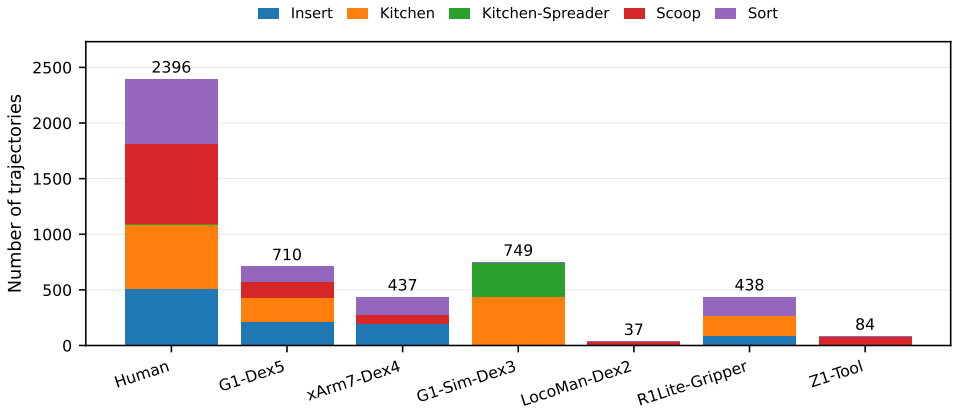

Figure 8: **Composition of the EgoX dataset across embodiments and tasks.** Each bar represents the total number of trajectories collected for one embodiment, with stacked segments indicating the corresponding task composition. Numbers above the bars denote the total number of trajectories for each embodiment.

Data composition. The dataset is deliberately heterogeneous in both embodiment and task composition. Human demonstrations constitute the largest source, with 2,396 trajectories (49.4% of the full dataset), while demonstrations collected on physical robots contribute 1,706 trajectories (35.2%). The remaining 749 trajectories (15.4%) are collected with the simulated G1 embodiment. Among the robot embodiments, G1-DEX5 and G1-SIM-DEX3 provide the largest individual datasets, with 710 and 749 trajectories, respectively, while the remaining platforms provide smaller but morphologically diverse sources of manipulation experience.

Importantly, the dataset is not constructed by replicating the same task set across every embodiment. Instead, different embodiments contribute different subsets of tasks according to their morphology and manipulation capabilities. Cross-embodiment overlap therefore arises through semantically corresponding observation and action modalities rather than requiring identical tasks or actuator-level action spaces across all platforms.

Table 3: **Task progress scoring used for evaluation.** Intermediate stages receive partial credit, while completing all stages yields a normalized task progress score of 1.

Task	**Stage scoring**	$\mathbf{S}_{\max}$
SCOOP	Grasp scoop: 1; approach target material: 1; scoop and lift material: 2; deposit into target bin: 1. For Z1-TOOL, the scoop is fixed to the arm and the grasp stage is omitted.	5 / 4 (Z1)
SORT	Three objects are evaluated independently. For each object: pick: 1; approach a bin: 3 if the bin color matches the object and 1 otherwise; place into a bin: 3 if the bin color matches the object and 1 otherwise.	21
INSERT	Approach the T-shaped block: 1; grasp the block: 2; insertion: 2 for successful insertion, 1 for a near-successful insertion, and 0 otherwise.	5
KITCHEN-BREAD	For R1LITE-GRIPPER and G1-SIM-DEX3: pick sauce container, locomote to workspace, place container, pick spreading tool, dip into sauce, spread onto bread, and place the tool back; each stage receives 1 point. For G1-DEX5, retrieval additionally includes opening the refrigerator, grasping the container, and closing the refrigerator.	7 / 9 (G1)
KITCHEN-CLEANUP	Grasp the tool: 2. Place the tool intothe container: 2.	4

A.7 TASK AND EVALUATION PROTOCOL

We evaluate each policy using a stage-based task progress score that provides partial credit for intermediate task completion. For an episode with accumulated stage score s and task-specific

maximum score $S_{\max}$, we report the normalized task progress score as $P = s/S_{\max}$. Table 3 summarizes the stage definitions used for each task.

A.8 Simulation Details

Simulation setup. We simulate `G1-Sim-Dex3`, a Unitree G1 humanoid with three-finger Dex3 hands, in a kitchen environment built in Isaac Sim and Isaac Lab. In `Kitchen-Bread`, the robot retrieves a sauce can from the refrigerator, carries it to the worktop, and spreads its contents onto bread; the shorter `Kitchen-Cleanup` task requires picking up the spreader and placing it in a container. Objects interact through rigid-body contact, while sauce transfer is rendered visually without fluid simulation. Demonstrations are collected by teleoperation with PICO controllers, and the egocentric head-camera and wrist-camera images are rendered offline with RTX. Physics runs at 200 Hz and the locomotion controller at 50 Hz.

Randomization. At reset, we randomize the robot and object positions, object sizes, aisle width, and worktop height, keep the spreader supported by the plate, and check each sampled layout for clearance. We also vary the bread shape and the positions of background objects and furniture. During offline rendering, we randomize object and room materials, lighting, camera poses and fields of view, and image properties such as color, noise, blur, and compression. These visual changes leave the recorded motions and action labels unchanged. Each rendered episode uses a fixed visual configuration, with noise varying across frames, and the dataset also retains variants close to the original appearance.

Evaluation. Each task is evaluated on 50 fixed trials: 32 ID and 18 OOD. OOD trials use real-world-inspired or cross-task appearances, and `Kitchen-Bread` OOD trials also include unseen plate–spreader layouts. Trial cases were refined through pilot evaluations and then fixed for all methods. `Kitchen-Bread` and `Kitchen-Cleanup` trials last 90 and 45 simulated seconds, respectively. Actions are executed at 30 Hz, with 45 actions executed before each new prediction. `Kitchen-Bread` success requires spreading sauce onto the bread and, at the end of the trial, a released spreader settled on a support. `Kitchen-Cleanup` success requires grasping the spreader and placing it inside the container, released and supported, at any point during the trial. We score nine prerequisite-dependent stages for `Kitchen-Bread` and four for `Kitchen-Cleanup`, plus one point for success, giving maximum scores of 10 and 5; normalizing by these maxima gives the task progress score reported in Sec. 3.

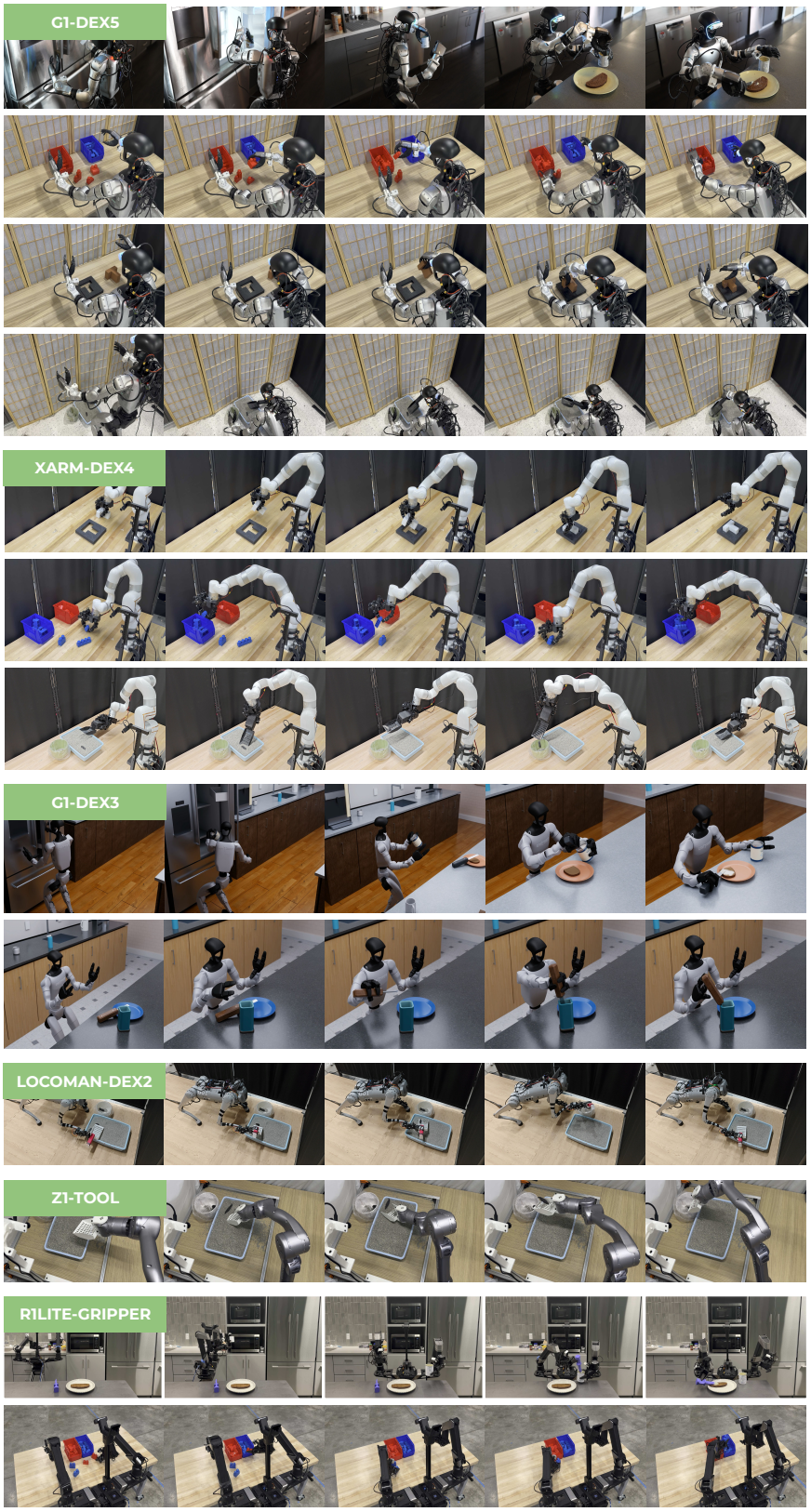

Figure 9: **Overview of tasks for each embodiment.** From top to bottom: G1-DEX5, XARM7-DEX4, G1-DEX3, LOCOMAN-DEX2. Z1-TOOL, and R1LITE-GRIPPER.