Egocentric Cross-Embodiment Manipulation with Embodiment Dreaming

Binghong Chen*, Yaru Niu*, Changwei Yao*, Zhenlong Fang*, Shangtao Li*
Revanth Krishna Senthilkumaran, Shuai Zhou, Yuemin Mao, Hao Zhang, Bingqing Chen
Chen Qiu, Eric H. Tseng, Changliu Liu, Jonathan Francis, Ding Zhao
*Equal contribution

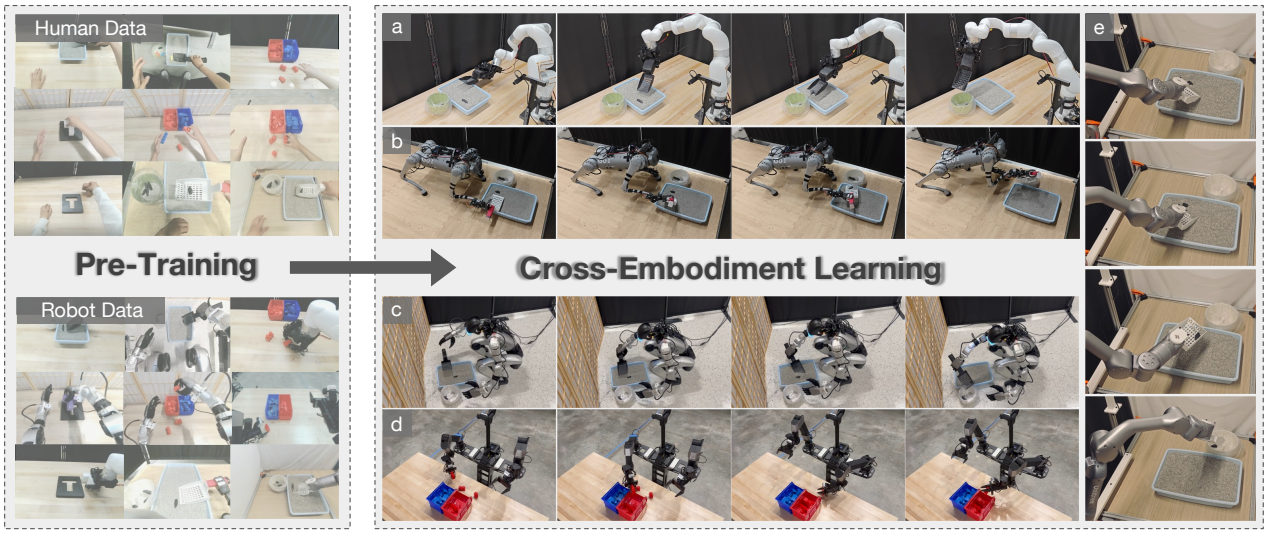

Fig. 1: EgoX represents human and robot demonstrations as unified egocentric observation-action trajectories organized by corresponding modality types. It pretrains a Modularized Cross-Embodiment Transformer with Embodiment Dreaming (MXT+) on mixed human-robot data (left), then finetunes the shared representation for target embodiments and manipulation tasks (right).

***Abstract*—Scaling robot manipulation requires learning from diverse embodied data, including demonstrations from both humans and heterogeneous robots. Such data can provide broad coverage of objects, scenes, tasks, and interaction strategies, but large embodiment gaps in sensing, kinematics, motion patterns, and action spaces make it difficult to train a single policy from the combined data. We introduce EgoX, an egocentric cross-embodiment learning framework for manipulation. EgoX records human and robot demonstrations as observation-action trajectories in a unified egocentric format, pretrains a Modularized Cross-Embodiment Transformer with Embodiment Dreaming (MXT+) on the resulting mixture, and finetunes it on target-embodiment demonstrations. MXT+ combines embodiment-specific tokenizers and action experts with a shared Transformer trunk, allowing heterogeneous observations and actions to update a common manipulation representation while preserving embodiment-specific sensing and control interfaces. Its embodiment-dreaming objective encourages the shared trunk to model each body's dynamics by predicting future embodiment latents from egocentric visual and proprioceptive features. These latent targets are produced with EMA copies of MXT+'s own modality encoders, regularizing the shared trunk without an additional encoder pretraining stage or any deployment-time computation. We evaluate EgoX on multiple manipulation tasks across five diverse embodiments spanning humanoid, wheeled mobile, tabletop arm, quadrupedal, and tool-use platforms with different end-effectors. Experiments on cross-embodiment pretraining and embodiment dreaming show improved downstream manipulation performance and transfer across substantially different embodiments.**

I. Introduction

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 [25, 32, 18, 9, 36], while egocentric human data provides another scalable source of physical experience [16, 17, 46]. However, combining these sources for low-level manipulation remains difficult because humans and robots differ in morphology, end-effectors, proprioception, viewpoints, action semantics, and motion patterns.

We introduce EgoX, an egocentric cross-embodiment framework that organizes human and robot demonstrations as unified observation-action trajectories and trains MXT+, a Modularized Cross-Embodiment Transformer with Embodiment Dreaming. EgoX treats humans and robots as embodiments of a shared manipulation process: rather than viewing each body as a monolithic interface, EgoX decomposes demonstrations into anatomically or functionally corresponding modality types, such as egocentric vision, body motion, end-effector motion, and local end-effector geometry. MXT+ uses embodiment-specific tokenizers and action experts around a shared Transformer trunk, allowing heterogeneous embodiments to update a common representation while preserving their own sensing and control interfaces. Its *embodiment*

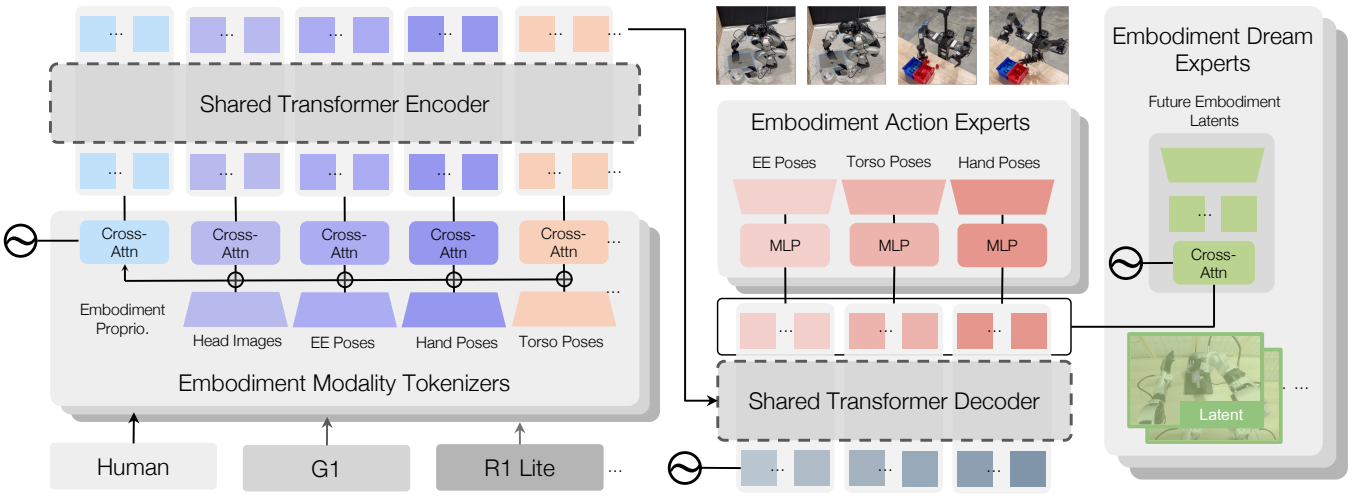

Fig. 2: **MXT+ architecture.** Embodiment-specific modality tokenizers map per-embodiment observations (egocentric images, end-effector, hand, and torso poses) into tokens that a shared Transformer encoder-decoder trunk fuses across embodiments. Embodiment action experts decode the trunk output into embodiment-specific action chunks, while embodiment-dreaming experts predict future embodiment latents supervised by EMA copies of the corresponding modality tokenizers.

dreaming objective predicts future embodiment latents from egocentric visual and proprioceptive features, encouraging the shared trunk to model body-specific dynamics without adding deployment-time computation.

We evaluate EgoX on `Sort` and `Scoop` across five embodiments: `G1`, `R1Lite`, `xArm7`, `LocoMan`, and `Z1`. Our experiments compare ACT, MXT+ without pretraining, and MXT+ with cross-embodiment pretraining, and ablate embodiment dreaming by comparing against MXT. The results show that cross-embodiment pretraining improves downstream performance, and that embodiment dreaming improves both performance and learned embodiment-latent representations. In summary, our contributions are:

1) EgoX, a unified egocentric human-robot demonstration format for cross-embodiment manipulation.
2) MXT+, a modular cross-embodiment Transformer with embodiment dreaming for future-latent prediction.
3) Real-robot experiments across two tasks and five embodiments showing the benefits of cross-embodiment pretraining and embodiment dreaming.

II. Methodology

EgoX has three stages: collect unified egocentric trajectories from humans and robots, pretrain MXT+ on mixed cross-embodiment demonstrations, and finetune the pretrained policy on target-embodiment tasks. As shown in Fig. 2, MXT+ shares a Transformer trunk across embodiments while using modular tokenizers, action experts, and embodiment dreaming to preserve embodiment-specific sensing, control, and dynamics.

A. Egocentric Cross-Embodiment Data Collection

EgoX stores human and robot demonstrations as egocentric observation-action trajectories in a shared dictionary format. The key is to decompose heterogeneous bodies into smaller modality types that correspond across embodiments, while keeping tokenization and control embodiment-specific. Observations include main-camera images, body and end-effector poses, and local end-effector geometry; actions use the corresponding embodiment-specific body, end-effector, and end-effector-link targets. All poses are expressed in a unified egocentric frame $\mathcal{F}_u$ attached to the main camera or camera-carrying body, and modality masks indicate which entries are valid for each embodiment:

$$\boldsymbol{o}_t^e = \{\boldsymbol{I}^e_{\text{main},t},\, \boldsymbol{x}^e_{\text{body},u,t},\, \boldsymbol{x}^e_{\text{L/R ee},u,t},\, \boldsymbol{p}^e_{\text{L/R ee link},t}, \dots\}, \tag{1}$$

$$\boldsymbol{a}_t^e = \{\boldsymbol{x}^{e,\text{target}}_{\text{body},u,t},\, \boldsymbol{x}^{e,\text{target}}_{\text{L/R ee},u,t},\, \boldsymbol{p}^{e,\text{target}}_{\text{L/R ee link},t}, \dots\}. \tag{2}$$

Here e may denote either a human or robot embodiment, and L/R end-effector entries define anatomical or functional modality types for bimanual, unimanual, hand, gripper, and tool embodiments. This format lets humans and heterogeneous robots share manipulation-relevant structure while preserving embodiment-specific sensing and control interfaces.

B. Modularized Cross-Embodiment Transformer

To learn from heterogeneous human and robot demonstrations, EgoX uses MXT+, a Modularized Cross-Embodiment Transformer with Embodiment Dreaming that builds on prior MXT-style modular tokenization and action decoding [24]. MXT+ places embodiment-specific tokenizers and action experts around a shared encoder-decoder Transformer trunk. For each valid observation modality $m \in \mathcal{M}^e_{\text{obs}}$, a tokenizer $T_{e,m}$ maps the raw observation into modality tokens, and the shared trunk fuses the concatenated tokens:

$$\begin{aligned} \boldsymbol{z}_t^e &= \text{concat}_{m \in \mathcal{M}^e_{\text{obs}}} T_{e,m}(\boldsymbol{o}_t^e[m])\,, \\ \boldsymbol{h}_t^e &= G_\theta(\boldsymbol{z}_t^e)\,. \end{aligned} \tag{3}$$

For each valid action modality $m \in \mathcal{M}^e_{\text{act}}$, an embodiment-specific MLP action expert predicts each step of the action chunk from the trunk output:

$$\hat{\boldsymbol{a}}^e_{t+\ell}[m] = D^{\text{MLP}}_{e,m,\ell}(\boldsymbol{h}_t^e)\,, \qquad \ell = 1, \dots, h. \tag{4}$$

Here h is the action horizon. This modular design lets embodiments share egocentric visual and manipulation structure through the trunk while keeping action dimensionality, masks, and control semantics embodiment-specific.

C. Embodiment Dreaming

Behavioral cloning alone supervises the policy through demonstrated actions, but it does not explicitly require the shared representation to model how each embodiment's state evolves after interaction with the environment. MXT+ therefore uses *embodiment dreaming*, a predictive auxiliary objective that asks the policy to infer a future *embodiment latent*: a compact representation of selected future observation modalities for the active embodiment. By predicting how the egocentric view, robot joint configuration, and end-effector joint configuration will evolve, this objective encourages the shared trunk to be embodiment-aware and to capture dynamics that differ across bodies, end-effectors, and control interfaces.

1) Embodiment latent.: For embodiment e, let $\mathcal{M}^e_{\text{dream}} \subseteq \mathcal{M}^e_{\text{obs}}$ denote the future observation modalities used as dreaming sources. In EgoX, these sources are restricted to egocentric image features, robot joint configurations, and end-effector joint configurations. Although robot and end-effector joint configurations are not used as cross-embodiment action targets, they provide valuable embodiment-specific state representations for learning body dynamics. MXT+ builds an embodiment tokenizer A_e that maps encoded modality features directly into an embodiment latent. Given future observations at offset k, an EMA teacher produces the target latent

$$\begin{aligned} \boldsymbol{F}^{e,\text{ema}}_{t+k} &= \text{concat}_{m\in\mathcal{M}^e_{\text{dream}}}\left[\phi^{\text{ema}}_{e,m}\left(\boldsymbol{o}^e_{t+k}[m]\right) + \boldsymbol{r}_{m,t+k}\right], \\ \boldsymbol{y}^{e,\star}_{t+k} &= \text{stopgrad}\left[A^{\text{ema}}_e\left(\boldsymbol{F}^{e,\text{ema}}_{t+k}\right)\right], \end{aligned} \tag{5}$$

where $\phi^{\text{ema}}_{e,m}$ is the EMA copy of the modality encoder, $\boldsymbol{r}_{m,t+k}$ denotes the positional feature added before aggregation, and A^{ema}_e is the EMA copy of the embodiment tokenizer. The embodiment tokenizer directly produces the latent supervised by the dreaming head, without pooling over multiple output tokens. The stop-gradient operation ensures that the dreaming target acts as a slowly evolving supervisory signal rather than a second optimization path.

2) Dreaming head.: An embodiment-dreaming expert reads from the shared trunk output and predicts one future embodiment latent for each offset in $\mathcal{K}$. The dream labels are sampled from future observations using the same dataset pipeline as action chunks, with padding masks for offsets beyond the episode horizon. We provide the full prediction form in Appendix IV-C.

3) Latent prediction loss.: We supervise the dreamed embodiment latents with a cosine alignment loss and an optional magnitude loss:

$$\begin{aligned} \mathcal{L}^e_{\text{dream}} = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\Big[1 - \cos\big(\hat{\boldsymbol{y}}^e_{t+k}, \boldsymbol{y}^{e,\star}_{t+k}\big) \\ + \beta\ell_\delta\left(\left\|\hat{\boldsymbol{y}}^e_{t+k}\right\|_2, \left\|\boldsymbol{y}^{e,\star}_{t+k}\right\|_2\right)\Big], \end{aligned} \tag{6}$$

where ℓ_δ is the smooth L1 loss and β weights magnitude alignment. Smooth L1 can also be used directly when the dream target is a raw normalized state rather than a latent. In practice, the latent form is useful because it asks the trunk to predict future embodied state at the level of features, including visual and joint-configuration dynamics, without reconstructing high-dimensional raw observations such as images.

4) EMA target update.: After each optimizer step, the EMA teacher is updated from the student modality tokenizers and embodiment tokenizer, providing a stable latent target while the student representation evolves. During deployment, the dreaming head and EMA teacher are removed or ignored; only the action experts are executed. The update equation is given in Appendix IV-C.

D. Cross-Embodiment Pretraining and Target Finetuning

EgoX first pretrains MXT+ on a mixture of human and robot embodiments using action prediction and embodiment dreaming, then finetunes the pretrained model on target-embodiment demonstrations. The shared trunk learns egocentric manipulation priors from all available data, while embodiment-specific tokenizers and experts preserve each robot's sensing and action statistics. We provide the full objectives and training procedure in Appendix IV-C and Algorithm 1.

1) Inference.: At deployment, MXT+ uses only the target embodiment's observation tokenizers, shared trunk, and applicable action experts. Using the egocentric observation dictionary, the policy predicts an action chunk in the modular action format and sends those targets to the embodiment's low-level controller. Embodiment-dreaming experts and EMA teachers are used only during training and discarded at inference.

III. Experiments

We conduct experiments to answer the following research questions: 1) Does egocentric cross-embodiment pretraining improve downstream dexterous manipulation compared to training from scratch on target-embodiment data? 2) Does embodiment dreaming learn useful representations and improve cross-embodiment manipulation performance?

A. Experimental Setup

We evaluate EgoX on two manipulation tasks, `Sort` (Building Block Sorting) and `Scoop` (Cat Litter Scooping), across five embodiments spanning humanoid, wheeled mobile, tabletop arm, quadrupedal, and tool-use platforms: `G1`, `R1Lite`, `xArm7`, `LocoMan`, and `Z1`. This setup also covers diverse end-effectors, including five-finger dexterous hands, four-finger dexterous hands, parallel-jaw grippers, loco-manipulator grippers, and fixed non-actuated tool end-effectors. For each task and embodiment, we evaluate both in-distribution and out-of-distribution settings and report success rate and task score. We compare ACT, MXT+ trained from scratch, and MXT+ with cross-embodiment pretraining, and ablate embodiment dreaming by comparing against MXT. For two-stage methods, we first pretrain on data from all embodiments while holding out the target task for the target embodiment, then finetune

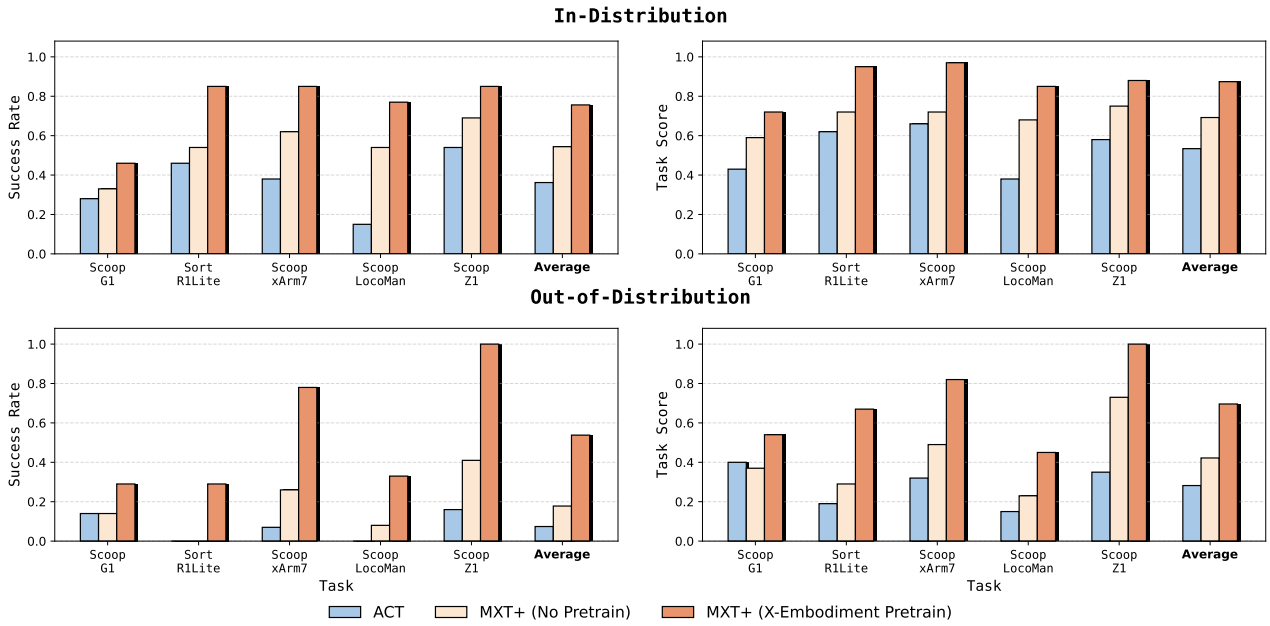

Fig. 3: **Main results.** We compare ACT, MXT+ without pretraining, and MXT+ with cross-embodiment pretraining across `Sort` and `Scoop` on five embodiments under in-distribution and out-of-distribution settings.

on that target task and embodiment. Full task definitions, embodiment specifications, evaluation protocol, and baseline details are provided in Appendix IV-E.

B. Egocentric Cross-Embodiment Pretraining Improves Downstream Performance

We first study whether egocentric cross-embodiment pretraining improves downstream policy learning compared to training directly on target-embodiment demonstrations. In Fig. 3, MXT+ with cross-embodiment pretraining improves both success rate and task score across nearly all task-embodiment pairs, with the clearest gains under out-of-distribution evaluation. In-distribution, pretraining raises the average performance over MXT+ trained from scratch and is especially helpful for lower-performing settings such as `Scoop` on `LocoMan`. Out-of-distribution, the gap becomes much larger: pretrained MXT+ substantially improves `Scoop` on `xArm7` and `Z1`, recovers nonzero performance on `Sort` with `R1Lite`, and gives the best average success rate and task score. This suggests that mixed human-robot demonstrations provide visual, geometric, and interaction patterns that are not covered by target-task data alone.

C. Embodiment Dreaming Improves Cross-Embodiment Learning

Finally, we compare MXT+ against MXT, its counterpart without the embodiment-dreaming objective. Embodiment dreaming asks the policy to predict future embodiment latents in addition to action chunks. As shown in Fig. 4, embodiment dreaming improves both metrics on all evaluated `Scoop` embodiments. The largest success-rate gain appears on `xArm7`, while `LocoMan` also benefits despite its different quadrupedal morphology and gripper-based loco-manipulator interface; `Z1` starts from a stronger baseline but still improves

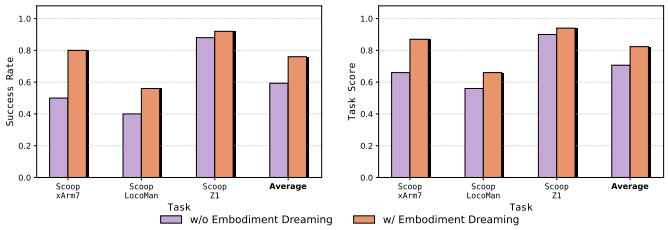

Fig. 4: **Embodiment dreaming ablation.** Predicting future embodiment latents improves success rate and task score on `Scoop` across `xArm7`, `LocoMan`, and `Z1`.

in both success rate and task score. These gains suggest that embodiment dreaming helps MXT+ learn not only an action policy, but also predictive latent structure for how each body, end-effector, and egocentric view evolves during contact-rich tool use. This is particularly useful for cross-embodiment learning, where the same task must transfer across different kinematics, camera viewpoints, and action spaces.

IV. Conclusions

We presented EgoX, an egocentric cross-embodiment framework that organizes human and robot demonstrations in a unified format and trains MXT+, a Modularized Cross-Embodiment Transformer with Embodiment Dreaming. MXT+ combines embodiment-specific tokenizers and action experts with a shared trunk, preserving embodiment-specific sensing and control interfaces while sharing task-relevant structure across bodies and end-effectors. Embodiment dreaming trains the shared trunk to predict future embodiment latents, encouraging embodiment-aware dynamics modeling without deployment-time computation. Experiments show that cross-embodiment pretraining improves downstream policy learning over target-only training and that embodiment dreaming further improves transfer.

References

[1] 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.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael 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. *arXiv preprint arXiv:2410.24164*, 2024.

[3] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. In *Robotics: Science and Systems (RSS)*, 2025. URL https://arxiv.org/abs/2505.06111. arXiv:2505.06111.

[4] Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-N-On: Scaling Egocentric Manipulation with in-the-wild and on-task Data. *arXiv preprint arXiv:2511.15704*, 2025. doi: 10.48550/arXiv.2511.15704. URL http://arxiv.org/abs/2511.15704.

[5] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large Video Planner Enables Generalizable Robot Control. *arXiv preprint arXiv:2512.15840*, 2025. doi: 10.48550/arXiv.2512.15840. URL http://arxiv.org/abs/2512.15840.

[6] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. *arXiv preprint arXiv:2402.19249*, 2024.

[7] Mingfei Chen, Yifan Wang, Zhengqin Li, Homanga Bharadhwaj, Yujin Chen, Chuan Qin, Ziyi Kou, Yuan Tian, Eric Whitmire, Rajinder Sodhi, Hrvoje Benko, Eli Shlizerman, and Yue Liu. Flowing from Reasoning to Motion: Learning 3D Hand Trajectory Prediction from Egocentric Human Interaction Videos. *arXiv preprint arXiv:2512.16907*, 2025. doi: 10.48550/arXiv.2512.16907. URL http://arxiv.org/abs/2512.16907.

[8] 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. *arXiv preprint arXiv:2410.08464*, 2024. URL https://arxiv.org/abs/2410.08464.

[9] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

[10] Yicheng Feng, Wanpeng Zhang, Ye Wang, Hao Luo, Haoqi Yuan, Sipeng Zheng, and Zongqing Lu. Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos. *arXiv preprint arXiv:2512.13080*, 2025. doi: 10.48550/arXiv.2512.13080. URL http://arxiv.org/abs/2512.13080.

[11] 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.

[12] 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 generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

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

[14] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

[15] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[16] 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)*, pages 13226–13233. IEEE, 2025.

[17] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of Human to Robot Transfer in Vision-Language-Action Models. *arXiv preprint arXiv:2512.22414*, 2025. URL https://arxiv.org/abs/2512.22414.

[18] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[19] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from In-the-wild Human Videos using Data-Editing. *arXiv preprint arXiv:2508.09976*, 2025. doi: 10.48550/arXiv.2508.09976. URL http://arxiv.org/abs/2508.09976.

[20] Guangrun Li, Yaoxu Lyu, Zhuoyang Liu, Chengkai Hou, Jieyu Zhang, and Shanghang Zhang. H2R: A Human-to-Robot Data Augmentation for Robot Pre-training from Videos. *arXiv preprint arXiv:2505.11920*, 2026. doi: 10.

48550/arXiv.2505.11920. URL http://arxiv.org/abs/2505.11920.

[21] Yangcen Liu, Woo Chul Shin, Yunhai Han, Zhenyang Chen, Harish Ravichandar, and Danfei Xu. ImMimic: Cross-Domain Imitation from Human Videos via Mapping and Interpolation. *arXiv preprint arXiv:2509.10952*, 2025. URL https://arxiv.org/abs/2509.10952.

[22] Hao Luo, Yicheng Feng, Haoqi Yuan, Wanpeng Zhang, Ye Wang, Sipeng Zheng, and Zongqing Lu. Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos. *arXiv preprint arXiv:2507.15597*, 2025. URL https://arxiv.org/abs/2507.15597.

[23] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[24] 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.

[25] 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: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[26] Younghyo Park, Jagdeep Singh Bhatia, Lars Ankile, and Pulkit Agrawal. Dexhub and dart: Towards internet scale robot data collection. *arXiv preprint arXiv:2411.02214*, 2024.

[27] 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. *arXiv preprint arXiv:2604.07607*, 2026.

[28] 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˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[29] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3352–3360. IEEE, 2025.

[30] 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.

[31] 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*, 2025.

[32] 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. *arXiv preprint arXiv:2405.12213*, 2024.

[33] Weikang Wan, Jiawei Fu, Xiaodi Yuan, Yifeng Zhu, and Hao Su. LodeStar: Long-horizon Dexterity via Synthetic Data Augmentation from Human Demonstrations. *arXiv preprint arXiv:2508.17547*, 2025. URL https://arxiv.org/abs/2508.17547.

[34] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-Horizon Imitation Learning by Watching Human Play. In *Proceedings of The 7th Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pages 201–221. PMLR, 2023. URL https://proceedings.mlr.press/v229/wang23a.html. arXiv:2302.12422.

[35] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation. In *Robotics: Science and Systems (RSS)*, 2024. URL https://arxiv.org/abs/2403.07788. arXiv:2403.07788.

[36] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *arXiv preprint arXiv:2409.20537*, 2024.

[37] Xiaomeng Xu, Jisang Park, Han Zhang, Eric Cousineau, Aditya Bhat, Jose Barreiros, Dian Wang, and Shuran Song. Hommi: Learning whole-body mobile manipulation from human demonstrations. *arXiv preprint arXiv:2603.03243*, 2026.

[38] Richard Yang, Ri-Zhao Qiu, Xuxin Cheng, 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.

[39] 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. *arXiv preprint arXiv:2410.11758*, 2024.

[40] Jessica Yin, Haozhi Qi, Youngsun Wi, Sayantan Kundu, Mike Lambeta, William Yang, Changhao Wang, Tingfan Wu, Jitendra Malik, and Tess Hellebrekers. OSMO: Open-Source Tactile Glove for Human-to-Robot Skill Transfer. *arXiv preprint arXiv:2512.08920*, 2025. doi: 10.48550/arXiv.2512.08920. URL http://arxiv.org/abs/2512.08920.

[41] Zhenhan Yin, Xuanhan Wang, Jiahao Jiang, Kaiyuan Deng, Pengqi Chen, Shuangle Li, Chong Liu, Xing Xu, Jingkuan Song, Lianli Gao, and Heng Tao Shen. MiVLA:

Towards Generalizable Vision-Language-Action Model with Human-Robot Mutual Imitation Pre-training. *arXiv preprint arXiv:2512.15411*, 2025. doi: 10.48550/arXiv.2512.15411. URL http://arxiv.org/abs/2512.15411.

[42] Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. EgoMI: Learning Active Vision and Whole-Body Manipulation from Egocentric Human Demonstrations. *arXiv preprint arXiv:2511.00153*, 2025. doi: 10.48550/arXiv.2511.00153. URL http://arxiv.org/abs/2511.00153.

[43] 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.

[44] Chi Zhang, Penglin Cai, Haoqi Yuan, Chaoyi Xu, and Zongqing Lu. UniTacHand: Unified Spatio-Tactile Representation for Human to Robotic Hand Skill Transfer. *arXiv preprint arXiv:2512.21233*, 2025. doi: 10.48550/arXiv.2512.21233. URL http://arxiv.org/abs/2512.21233.

[45] 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.

[46] 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.

[47] 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. *arXiv preprint arXiv:2509.04443*, 2025. URL https://arxiv.org/abs/2509.04443.

APPENDIX

A. Egocentric Cross-Embodiment Data Collection Details

EgoX collects demonstrations from a set of embodiments $\mathcal{E}$, including humans and robot embodiments, with various end-effectors, sensors, and control interfaces. For each embodiment $e \in \mathcal{E}$, an episode is stored as a sequence

$$\mathcal{D}_e = \{(\boldsymbol{o}_t^e, \boldsymbol{a}_t^e)\}_{t=1}^{T}, \tag{7}$$

where $\boldsymbol{o}_t^e$ is a dictionary of observations and $\boldsymbol{a}_t^e$ is a dictionary of embodiment-specific action targets. Across embodiments, observations include egocentric main-camera images, body and end-effector pose proprioception, and end-effector joint pose proprioception, with optional embodiment-specific streams such as wrist-camera images, robot joint states, tactile readings, and force feedback. Action targets are represented in the same modular format, including body pose commands, end-effector pose commands, end-effector joint pose commands, and, when available, velocity commands.

a) Unified embodiment-centric frame: To reduce avoidable coordinate mismatch across embodiments and data collection interfaces, EgoX expresses poses in a unified frame $\mathcal{F}_u$ attached to the egocentric main camera frame or the rigid body carrying the main camera. 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 being stored. This convention is used for both human demonstrations and robot demonstrations, so the dataset retains a consistent egocentric spatial interpretation even when data are collected through different interfaces and the downstream action spaces differ substantially.

b) Cross-embodiment demonstrations: EgoX stores human and robot demonstrations in a shared format of images, proprioceptive states, actions, masks, and embodiment metadata. Each demonstration is represented as a cross-embodiment trajectory with the same observation-action dictionary structure:

$$\boldsymbol{o}_t^e := \{\boldsymbol{I}^e_{\text{main},t},\, \boldsymbol{x}^e_{\text{body},u,t},\, \boldsymbol{x}^e_{\text{L/R ee},u,t},\, \boldsymbol{p}^e_{\text{L/R ee link},t}, \ldots\}, \tag{8}$$

$$\boldsymbol{a}_t^e := \{\boldsymbol{x}^{e,\text{target}}_{\text{body},u,t},\, \boldsymbol{x}^{e,\text{target}}_{\text{L/R ee},u,t},\, \boldsymbol{p}^{e,\text{target}}_{\text{L/R ee link},t}, \ldots\}. \tag{9}$$

Here e may denote either a human or robot embodiment. The L/R end-effector entries define anatomical or functional action modalities; for single-arm embodiments, the valid side is indicated by the modality masks. The terms $\boldsymbol{p}^e_{\text{L/R ee link},t}$ denote task-relevant 3D keypoints on each end effector, expressed in the corresponding end-effector local frame; examples include human fingertip positions, robot-hand fingertip positions, gripper contact-point positions, and tool contact-point positions. These quantities describe end-effector geometry rather than actuator joint coordinates. In this unified format, both humans and robots provide egocentric manipulation trajectories that pair visual observations with body, end-effector, and hand, gripper, or tool motion.

c) Masks and missing modalities: Not every embodiment provides every modality or action dimension. EgoX therefore associates each trajectory with modality masks for images, proprioception, and actions. During training, these masks indicate which observations and action targets are valid for a given embodiment, so unavailable modalities are ignored in tokenization and loss computation. This masking mechanism lets EgoX co-train on heterogeneous human and robot demonstrations while preserving semantically corresponding modality structure and embodiment-specific interfaces.

B. MXT+ Architecture Details

To learn from heterogeneous human and robot demonstrations, EgoX uses MXT+, a Modularized Cross-Embodiment Transformer with Embodiment Dreaming that builds on prior MXT-style modular tokenization and action decoding [24]. MXT+ separates embodiment-specific sensing and control interfaces from a shared policy backbone. It maps each embodiment's observations into modality tokens, fuses them with a shared encoder-decoder Transformer trunk, and decodes both action chunks and future embodiment latents through modular experts. Let an embodiment-specific observation be a dictionary

$$\boldsymbol{o}_t^e = \{\boldsymbol{o}_t^e[m]\}_{m \in \mathcal{M}^e_{\text{obs}}}, \tag{10}$$

where $\mathcal{M}^e_{\text{obs}}$ is the set of observation modalities for embodiment e.

a) Modality tokenizers: For each embodiment e and observation modality m, a tokenizer $T_{e,m}$ maps raw observations to a fixed number of tokens. The modalities are defined at an anatomical or functional level, such as egocentric vision, body state, end-effector pose, and local end-effector-link pose. This lets embodiments with different overall morphologies still expose compatible interfaces when they share a modality. Image modalities are encoded by a visual backbone and compressed by cross-attention aggregation, while state-like modalities are encoded by lightweight MLPs before cross-attention aggregation. The output tokens are concatenated:

$$\boldsymbol{z}_t^e = \text{concat}_{m \in \mathcal{M}^e_{\text{obs}}} T_{e,m}(\boldsymbol{o}_t^e[m]). \tag{11}$$

This decomposition preserves the semantic identity of each modality and avoids mixing heterogeneous signals too early. It also gives the model a stable interface for sharing structure across embodiments: the trunk always consumes modality tokens, while the tokenizers absorb embodiment-specific sensing differences such as dimensionality, camera availability, and kinematic parameterization.

b) Shared Transformer trunk: The shared trunk G_θ is an encoder-decoder Transformer. It receives the concatenated observation tokens and produces a fixed set of output tokens:

$$\boldsymbol{h}_t^e = G_\theta(\boldsymbol{z}_t^e). \tag{12}$$

The trunk parameters are shared across all embodiments during pretraining. Because modality tokenizers expose heterogeneous demonstrations through a common token interface, the trunk can learn reusable manipulation structure from

all embodiments without requiring their raw observations or actions to have the same dimensionality. This is the main conduit for cross-embodiment transfer: diverse embodiments update the same latent dynamics and decision-making module, while embodiment-specific tokenizers and experts preserve the differences needed for execution.

c) Modular action experts: For each valid action modality $m \in \mathcal{M}^e_{\text{act}}$, an embodiment-specific MLP action expert predicts each step of the action chunk from the trunk output:

$$\hat{\boldsymbol{a}}^e_{t+\ell}[m] = D^{\text{MLP}}_{e,m,\ell}(\boldsymbol{h}^e_t), \qquad \ell = 1, \dots, h. \tag{13}$$

Here h is the action horizon. This modular design is important for cross-embodiment learning: although two embodiments may differ substantially as complete systems, they often contain anatomically or functionally corresponding action modality types, such as body motion, end-effector motion, and local end-effector-link or hand motion. For example, the end-effector pose of a single-arm embodiment can correspond to the right end-effector pose of a bimanual embodiment, even though their tokenizers, action experts, and full action spaces are embodiment-specific. By assigning each modality its own action expert, MXT+ can reuse shared structure across embodiments while keeping modality dimensionality, control semantics, and validity masks embodiment-specific.

C. Training Objective Details

a) Dreaming head: An embodiment-dreaming expert Q_e reads from the shared trunk output and predicts a sequence of future embodiment latents:

$$\hat{\boldsymbol{Y}}^e_t = Q_e(\boldsymbol{h}^e_t) = \left\{\hat{\boldsymbol{y}}^e_{t+k}\right\}_{k\in\mathcal{K}}, \tag{14}$$

where $\mathcal{K}$ is a set of future offsets.

b) EMA target update: After each optimizer step, the teacher parameters are updated as

$$\theta^T_e \leftarrow \alpha\theta^T_e + (1-\alpha)\theta_e, \qquad \alpha \in (0,1), \tag{15}$$

where θ_e denotes the student parameters of the modality tokenizers and embodiment tokenizer used to construct the dream target.

c) Pretraining and finetuning objectives: Let $\mathcal{E}_{\text{pre}}$ be the set of embodiments used for pretraining, including humans and multiple robots. EgoX optimizes the sum of action prediction and embodiment-dreaming losses:

$$\mathcal{L}_{\text{pre}} = \sum_{e\in\mathcal{E}_{\text{pre}}} \left(\sum_{m\in\mathcal{M}^e_{\text{act}}} \lambda^e_m \mathcal{L}^e_{\text{act},m} + \lambda^e_{\text{dream}} \mathcal{L}^e_{\text{dream}} \right). \tag{16}$$

The action loss for modality m is the chunked behavioral cloning loss

$$\mathcal{L}^e_{\text{act},m} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}\sum_{\ell=1}^{h} \ell_1\left(\boldsymbol{a}^e_{j,\ell}[m], \hat{\boldsymbol{a}}^e_{j,\ell}[m]\right), \tag{17}$$

computed only on valid action dimensions according to the action mask. Given a target embodiment $e^\star$ and target task dataset $\mathcal{D}_{e^\star,\text{task}}$, EgoX initializes from the pretrained checkpoint and continues training with

$$\mathcal{L}_{\text{ft}} = \sum_{m\in\mathcal{M}^{e^\star}_{\text{act}}} \lambda^{e^\star}_m \mathcal{L}^{e^\star}_{\text{act},m} + \lambda^{e^\star}_{\text{dream}} \mathcal{L}^{e^\star}_{\text{dream}}. \tag{18}$$

D. Training Algorithm

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

Algorithm 1 EgoX pretraining and finetuning with embodiment dreaming

Input: Cross-embodiment datasets $\{\mathcal{D}_e\}_{e\in\mathcal{E}_{\text{pre}}}$, target dataset $\mathcal{D}_{e^\star,\text{task}}$
Input: Action horizon h, dream offsets $\mathcal{K}$, EMA decay α
Output: Target embodiment policy $\pi_{e^\star}$

Initialize MXT+ with modular tokenizers, shared trunk, action experts, and embodiment-dreaming experts
Initialize EMA dream teachers from the corresponding student tokenizers
for pretraining step $= 1, 2, \dots$ **do**
  Sample embodiment batches from $\{\mathcal{D}_e\}_{e\in\mathcal{E}_{\text{pre}}}$
  Predict action chunks and future embodiment latents
  Compute $\mathcal{L}_{\text{pre}}$ using Eq. (16)
  Update student parameters by gradient descent
  Update EMA teachers using Eq. (15)
end for
Initialize target policy $\pi_{e^\star}$ from the pretrained checkpoint
for finetuning step $= 1, 2, \dots$ **do**
  Sample target embodiment batch from $\mathcal{D}_{e^\star,\text{task}}$
  Predict target action chunks and, when enabled, target embodiment latents
  Compute $\mathcal{L}_{\text{ft}}$ using Eq. (18)
  Update student parameters and EMA teachers
end for
return $\pi_{e^\star}$ without executing dream heads at deployment

E. Experimental Setup

a) Tasks.: We evaluate EgoX on two manipulation tasks, with task labels used consistently with the figure annotations. In `Sort` (Building Block Sorting), the robot sorts blocks on a tabletop into containers of the corresponding colors; the blocks are randomly initialized within a rectangular region with varying poses. In `Scoop` (Cat Litter Scooping), the robot grasps or uses a scoop, collects cat litter randomly positioned inside a litter box, dumps it into a bin, and places the scoop back on the edge of the litter box. These tasks require different manipulation capabilities, such as semantic object selection, tool use, and contact-rich interaction.

b) Embodiments: To evaluate cross-embodiment learning, we deploy policies on five embodiments with distinct mobility, handedness, and end-effector designs: `G1`, a Unitree G1 humanoid with Inspire F1 hands, providing a bimanual legged/mobile embodiment with five-fingered dexterous hands;

`R1Lite`, a Galaxea R1 Lite wheeled bimanual mobile manipulator with parallel-jaw grippers, providing a wheeled mobile bimanual embodiment; `xArm7`, a tabletop xArm7 equipped with a Leap Hand, providing a unimanual embodiment with a four-fingered dexterous hand; `LocoMan`, a quadrupedal embodiment with loco-manipulators, i.e., grippers mounted on the front legs, evaluated in unimanual mode; and `Z1`, a tabletop Unitree Z1 arm with a fixed, non-actuated tool end-effector.

c) Evaluation settings and metrics: For each task and embodiment, we evaluate policies in both in-distribution and out-of-distribution settings. The in-distribution setting uses objects, initial configurations, and environmental conditions similar to those seen during target-embodiment training. The out-of-distribution setting uses held-out variations, such as novel object instances, appearances, poses, or scene configurations, to evaluate generalization beyond the target demonstration distribution. We report success rate and task score. Success rate measures the fraction of trials in which the policy completes the full task. Task score measures fine-grained task progress by assigning credit to intermediate substeps and normalizing the total score, which provides a more informative metric for long-horizon tasks where policies may partially complete the task.

d) Baselines and variants: We compare three policy variants. ACT is a strong imitation learning baseline trained directly on target-embodiment demonstrations, but it does not use MXT+'s modular tokenizers, action experts, or embodiment-dreaming experts. MXT+ without pretraining uses the same modular tokenizers, shared trunk, action experts, and embodiment-dreaming objective, but is trained from scratch on the target embodiment and task. MXT+ with cross-embodiment pretraining uses a two-stage protocol: first, it is pretrained on data from all embodiments while holding out the target task for the target embodiment; second, it is finetuned on that target task and embodiment. For ablations, we compare against MXT, which removes the embodiment-dreaming objective.

F. Related Work

Robot Learning from Human Data: Egocentric human data is a promising way to scale robot learning because it covers broad objects, scenes, tasks, and manipulation strategies beyond what can be collected through robot teleoperation alone. Early work uses human videos to learn transferable visual or interaction representations, including egocentric visual pretraining [23], hand-object and affordance-centric pretraining [30, 29], and latent-action objectives that extract action-like structure from videos without robot action labels [39, 3, 45]. Other methods strengthen the supervision extracted from human videos through visual-physical alignment [10], 3D hand trajectory prediction [7], video planning and world-model objectives [13, 5], or data editing and augmentation that make human videos more robot-compatible [19, 20, 33]. More recent work moves beyond using human data only as a representation source and treats humans as another embodiment to be co-trained with robots. This includes long-horizon guidance from human play [34], egocentric human-robot co-training [16, 17], explicit alignment or mutual imitation objectives [21, 41, 43], and large-scale egocentric VLA systems for humanoid, dexterous, mobile, and whole-body manipulation [38, 22, 11, 4, 28, 47, 42, 46, 37]. In parallel, data collection systems and datasets such as DexCap, ARCap, EgoDex, DexHub, EgoVerse, and tactile capture platforms improve the scale, quality, and modality coverage of human supervision [35, 8, 14, 26, 27, 40, 44]. EgoX builds on this direction by collecting structured egocentric demonstrations from both humans and robots and using them as cross-embodiment observation-action trajectories rather than only as visual pretraining data or hand-crafted transfer labels.

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 such as Open X-Embodiment, Octo, OpenVLA, π_0, $\pi_{0.5}$, Gemini Robotics, and GR00T demonstrate the value of scaling policy learning across tasks and robot embodiments [25, 32, 18, 2, 15, 31, 1]. A central challenge is that embodiment gaps can be large: robots may differ in camera setups, proprioceptive observations, kinematics, action dimensions, control frequencies, end-effectors, and motion patterns. Recent architectures address this heterogeneity by tokenizing variable observations, using flexible action readouts, or sharing a reusable policy trunk across embodiments. CrossFormer trains one Transformer policy over diverse robots by representing variable observations and actions as token sequences [9]; HPT learns a shared trunk with embodiment-specific tokenizers and action heads for heterogeneous visual and proprioceptive inputs [36]; and related methods study embodiment transfer through visual adaptation or cross-painting [6]. EgoX adopts this modular cross-embodiment perspective, but extends it to a unified human-robot setting: MXT+ uses modality-specific tokenizers, a shared Transformer trunk, modular action experts, and embodiment-dreaming experts so that human and robot demonstrations can update a common representation while preserving embodiment-specific sensing and control interfaces.

Embodiment Modeling and Predictive Supervision: Predictive objectives provide another way to learn representations for contact-rich manipulation by asking models to anticipate how observations evolve under actions. World-model approaches learn action-conditioned future dynamics from human or robot videos, including dexterous hand-object interaction models and large-scale robot world models trained from egocentric human videos [13, 12]. Latent-action and motion-aware VLA methods similarly learn compact action or dynamics tokens from video transitions, allowing policies to exploit unlabeled or heterogeneous videos without requiring every source to share the same robot action space [39, 3, 45, 11]. These works show that future prediction and compact latent dynamics can expose useful physical structure beyond static visual representation learning, but often require a separate representation, latent-action, or world-model training stage. EgoX brings this idea into cross-embodiment imitation learning through embod-

iment dreaming: MXT+ predicts future embodiment latents from egocentric visual features, robot joint configurations, and end-effector joint configurations using EMA copies of its own modality encoders, avoiding an extra stage for training the target encoder. The objective helps the shared trunk learn embodiment-aware dynamics while leaving the deployment-time action policy unchanged.