EGOWAM: World Action Models Beyond Pixels with In-the-Wild Egocentric Human Data

Baoyu Li*, Xinchen Yin*, Mengying Lin, Yixing Zhang, Danfei Xu
Georgia Institute of Technology

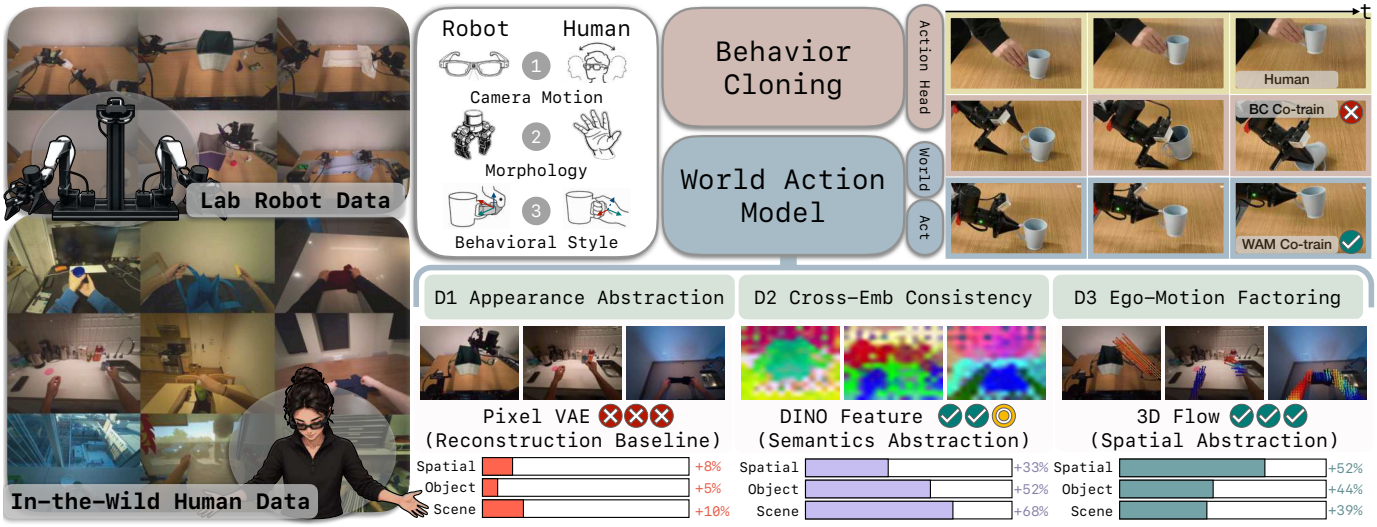


Fig. 1: **EGOWAM** co-trains on lab robot data and *in-the-wild egocentric human data*, separated by an embodiment gap of camera motion, morphology, and behavioral style. **(1)** BC co-training leaks this gap through the *sole* action head as inexecutable motions that harm performance; WAM co-training adds a channel predicting scene evolution, transferring human data through dynamics where actions cannot. **(2)** EGOWAM studies *what world representation* best enables transfer, comparing pixel VAE, DINO features, and 3D flow against three desiderata: appearance abstraction, cross-embodiment consistency, and ego-motion factoring.

***Abstract*—Egocentric human data offers scalable supervision for robot manipulation. However, behavior cloning entangles transferable content like objects, scenes, and task semantics, with non-transferable factors like human morphology, head motion, and behavioral style. We study whether World Action Models (WAMs) provide a better training signal by requiring policies to predict not only actions, but also how the scene evolves. The central question is what *world representation* best enables human-to-robot transfer. We hypothesize that an effective world target should abstract appearance, capture agent-invariant physical effects, and separate camera motion from environment change. We introduce EGOWAM, a controlled human-robot co-training framework that fixes the policy backbone, action head, and data mixture while varying only the world prediction target, comparing Pixel, DINO, and 3D motion flow. Across three real-world bimanual tasks, WAM co-training scales more effectively with in-the-wild egocentric human data than behavior cloning. Pixel-based prediction transfers weakly, while DINO and 3D flow yield substantial gains: DINO improves out-of-distribution object and scene generalization by up to 4×, and 3D flow improves in-domain performance by 20–30%.**

I. INTRODUCTION

Egocentric human data offers a promising source for scaling robot manipulation, providing diversity in objects, scenes, and behaviors that is prohibitively expensive to collect on robots [17, 12, 22]. A growing body of work exploits this through behavior-cloning (BC) co-training, retargeting human demonstrations into a robot-compatible action space and jointly training an imitation policy [21, 43]. Recent efforts [63, 42] show this paradigm can scale, but only with the human data that is carefully aligned to the robot in viewpoint, motion speed, and behavioral style. Without that bridge, action-level co-training injects human-like motions that the robot cannot execute and *degrades* downstream performance (Fig. 1). This is the *bitter lesson* of action-level co-training: the shared action decoder is the *sole* channel through which human data reaches the policy, forcing it to entangle transferable content (objects, scenes, semantics) with non-transferable execution (morphology, behavioral style), and the embodiment gap in the latter blocks transfer of the former.

World Action Models (WAMs) [54, 23, 28, 66] open a second supervision channel: an auxiliary world-model head predicts future states from a shared backbone, grounding action prediction in task-relevant dynamics. Because this channel operates on observations rather than actions, it is largely indifferent to the morphology and behavioral style that vary across embodiments. We hypothesize that task-relevant dynamics transfer across embodiments more readily than actions, so human data can shape the shared backbone through the world-model channel even when its action labels cannot. By

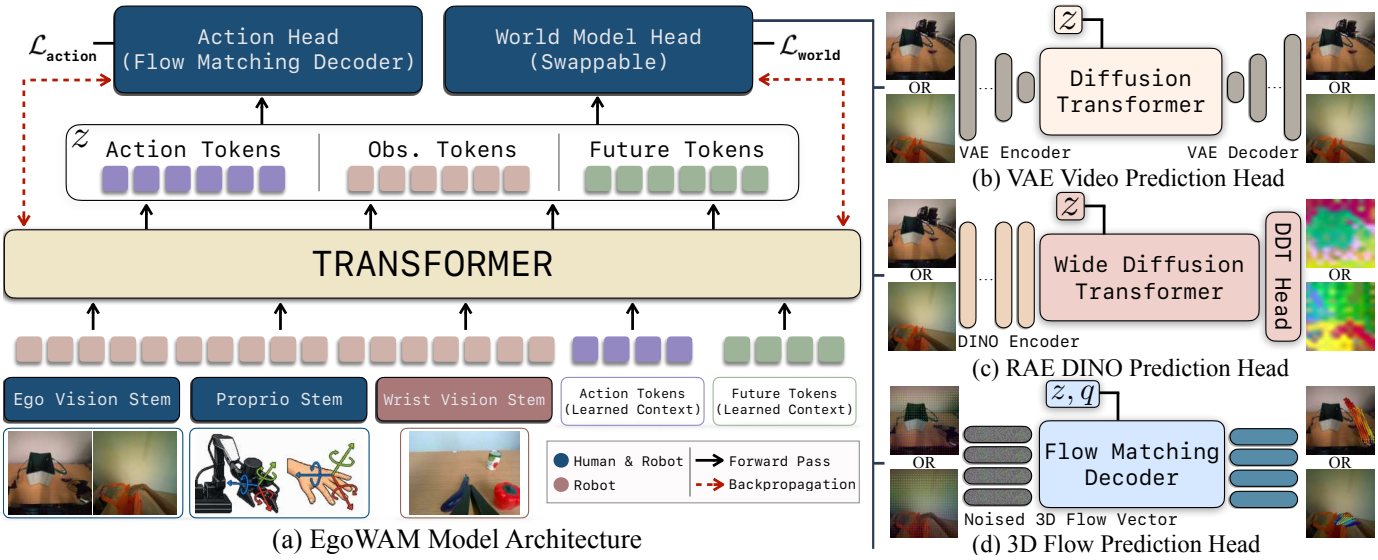


Fig. 2: **Model Architecture. (a)** EGOWAM builds on a Heterogeneous Pretrained Transformer with modality-specific stems (ego vision, proprioception, wrist vision), learned action and future tokens, a flow-matching action head, and a swappable world-model head supplying dynamics supervision that carries human data where actions cannot. The head supports three world targets: **(b)** a VAE head decoding pixel latents, **(c)** an RAE DINO-feature head, and **(d)** a 3D-flow head denoising camera-stabilized motion flow.

decoupling supervision into "*how the world evolves*" and "*how each embodiment acts*", WAMs induce a more shareable representation and enable more effective scaling from *large-scale in-the-wild human data* than BC co-training can achieve.

The promise of WAMs for human-robot co-training, however, hinges on a question that has not been systematically studied: ***what world representation enables effective transfer across embodiments under WAM co-training?*** Most WAMs predict pixels through a pretrained video VAE [54, 23, 28, 66] whose latent is optimized for photometric reconstruction and entangles motion with appearance, which we identify as the dominant failure mode of pixel-level WAM co-training. We argue that the world representation should satisfy three desiderata: *appearance abstraction*, *cross-embodiment consistency*, and *ego-motion factoring*. We then study two alternatives that satisfy them to different degrees: DINO features [38, 62], whose semantic prior abstracts appearance and aligns predictions across embodiments but remains spatially indexed on the image grid, and 3D motion flow [51, 32], which satisfies all three by construction through camera-frame-aligned geometric grounding.

We introduce **EGOWAM**, a framework for human-robot WAM co-training built upon a Heterogeneous Pretrained Transformer backbone [47, 42]. EGOWAM features a shared action head and a *swappable* world-model head trained jointly on human and robot data, isolating the effect of world representation under a matched backbone and data mixture. Evaluated on three real-world bimanual tasks spanning spatial, object, and scene generalization, our study makes three contributions:

- **A framework for controlled study of WAM co-training.** EGOWAM provides a single backbone with a swappable world-model head, enabling apples-to-apples comparison of world representations under identical action supervision and data mixtures.
- **Evidence that WAM co-training unlocks human-data scale.** Across in-domain and OOD evaluations, WAM co-training scales and transfers more consistently from large-scale natural egocentric human data than BC co-training, confirming that future-dynamics supervision is the missing channel where action-only supervision saturates.
- **World representation as the next critical axis.** Both DINO features and 3D motion flow substantially outperform the pixel-VAE baseline, with complementary strengths: DINO's semantic prior yields the strongest object and scene generalization, while 3D flow's geometric grounding yields the strongest spatial generalization and the highest overall scores.

II. HUMAN-ROBOT WAM CO-TRAINING

Problem setup. We consider an egocentric human dataset $\mathcal{D}_H = \{(o_t^H \;\; a_t^H)\}$ collected with Project Aria glasses [9] and a bimanual teleoperated robot dataset $\mathcal{D}_R = \{(o_t^R \;\; a_t^R)\}$. Both share an egocentric RGB stream I^{ego}; the robot additionally has wrist-mounted views I^{wrist}. Our goal is to co-train a single policy on $\mathcal{D}_H \cup \mathcal{D}_R$ such that in-the-wild human diversity improves robot performance without requiring human demonstrations to be executable on the robot.

Aligned action channel. We start from the strongest action-aligned BC baseline rather than a strawman [42]. Human and robot demonstrations are unified into a 14-D end-effector action space (per-arm 6-DoF $\mathrm{SE}(3)$ pose plus gripper). Robot actions are computed from joint angles via forward kinematics and expressed in the static ego-camera frame. Human hand poses, natively in the moving device frame T_t^{device}, are re-expressed in the *instantaneous* device frame at time t to factor out head ego-motion. Two residual mismatches remain: *(i)* speed: humans act faster than teleoperated robots, so we use embodiment-specific horizons (T_H=1 s, T_R=1 5 s) discretized into k steps; and *(ii)* workspace range: quantile normalization maps each action dimension's 1st/99th percentiles to $[-1 \;\; 1]$.

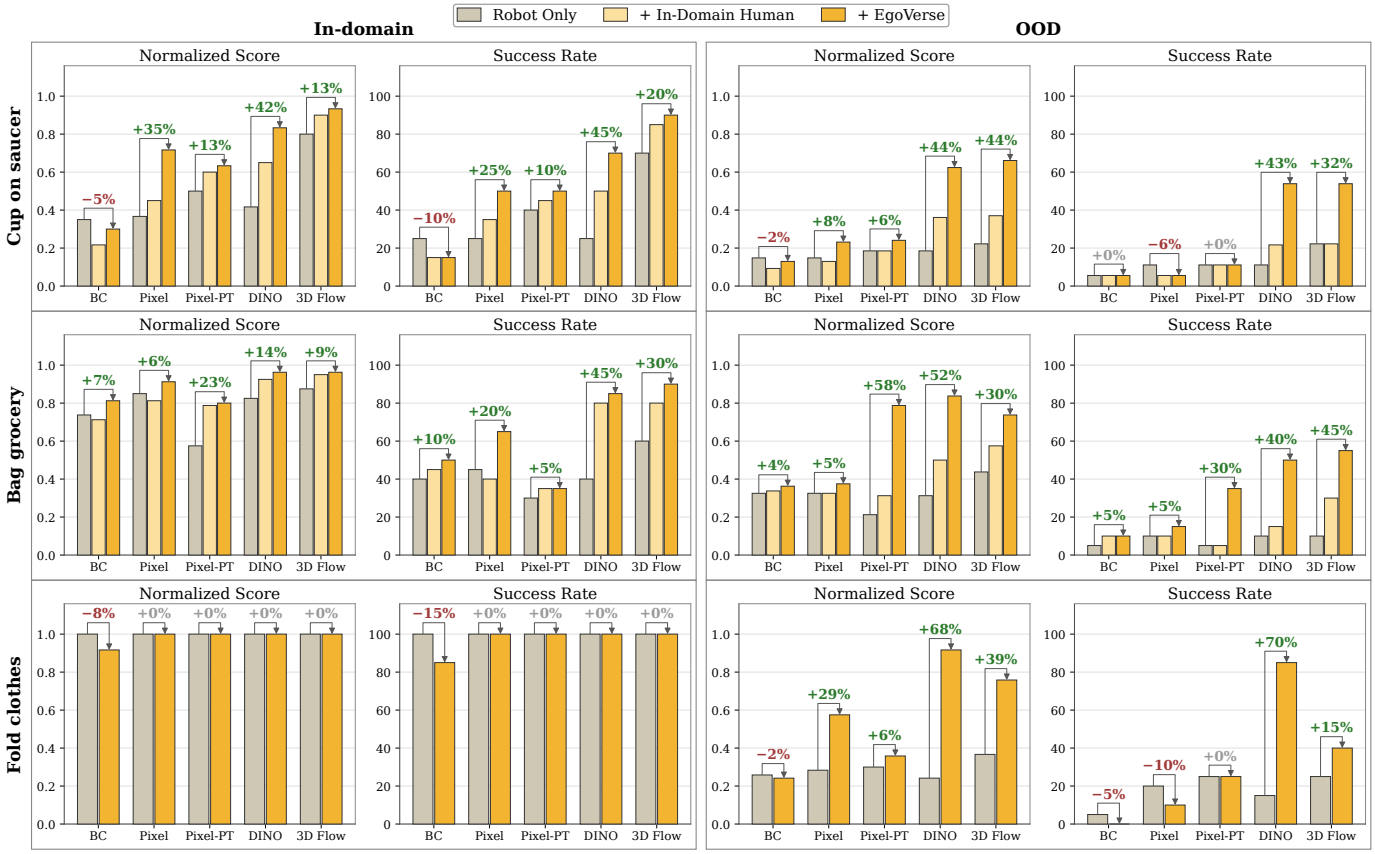


Fig. 3: **Quantitative Comparison on Real-World Rollouts.** WAM co-training consistently outperforms BC: human data often degrades BC yet yields large WAM gains. Pixel transfers weakly, while DINO drives the strongest OOD generalization and 3D Flow the largest ID spatial gains.

A shared encoder f_ϕ and action decoder π_θ are co-trained with

$$\mathcal{L}_{\text{BC-cotrain}}(\phi,\theta) = \sum_{\mathcal{D}\in\{\mathcal{D}_H,\mathcal{D}_R\}} \mathbb{E}_{(o,a)\sim\mathcal{D}}\, \mathcal{L}_{\text{BC}}(\pi_\theta(a \mid f_\phi(o)),\, a)\,, \tag{1}$$

realized as conditional flow-matching losses on both embodiments (App. B). Even with this alignment, action-level transfer leaves a gap: embodiment mismatch still leaks through the shared decoder as inexecutable motions (Fig. 1).

WAM as a second transfer interface. A World Action Model adds an auxiliary world-model head g_ψ that predicts a future state s_{t+T} from the same latent $z_t = f_\phi(o_t)$:

$$\mathcal{L}_{\text{WAM}} = \mathcal{L}_{\text{action}}(a_{t:t+k}) + \lambda\, \mathcal{L}_{\text{world}}(s_{t+T}) \tag{2}$$

Because $\mathcal{L}_{\text{world}}$ supervises the *shared trunk* through dynamics from *both* embodiments, human data can shape z_t through future-scene prediction even when action labels do not transfer faithfully, as shown in Fig. 2(a).

III. EgoWAM: A Controlled Study of World Representations

Desiderata. For WAM co-training to transfer from egocentric human data, we posit three desiderata for the world representation: **(D1) Appearance abstraction**: avoid rewarding photometric reconstruction, which forces the trunk to encode embodiment-specific appearance; **(D2) Cross-embodiment consistency**: represent the *effect*, not the *agent*, so a hand and a gripper induce similar physical change given similar action conditioning; **(D3) Ego-motion factoring**: targets should be invariant to camera motion, bridging head-worn human views and the robot's fixed ego-camera. The central question is ***what should the world target*** s_{t+T} ***be?***

Single-backbone instrument. EgoWAM builds on the HPT backbone [47, 42]: embodiment-specific stems tokenize each input into a shared trunk (Fig. 2). A ResNet-18 ego-vision stem encodes I_t^{ego}; a matching wrist stem encodes I_t^{wrist} on robot batches; a per-embodiment MLP encodes proprioception. Shared learnable *action* and *future* tokens query the trunk and route to two read-outs: a flow-matching action head π_θ that denoises action chunks $a_{t:t+k}$, and a *swappable* world-model head g_ψ that predicts s_{t+T} at the embodiment-specific horizon. Holding the trunk, action head, and data mixture fixed makes this a *controlled* study of world representation.

World-target instantiations. We instantiate three targets along the desiderata axis (Fig. 1), each trained under a shared linear path $s^\tau = (1-\tau)\epsilon + \tau s$ with noise $\epsilon \sim \mathcal{N}(0\ I)$: (1) **Pixel VAE** [46] predicts future ego frames in a pretrained video-VAE latent optimized for photometric reconstruction: $s = \text{VAE}(I_{t+T}^{\text{ego}})$. A diffusion transformer head predicts noise in this space. This target violates (D1)–(D3) and serves as our reconstruction baseline. (2) **DINO features** [38, 62] replace pixels with semantic patch features $s = \text{DINO}(I_{t+T}^{\text{ego}})$, predicted by a wide-DDT head in the RAE style. DINO satisfies

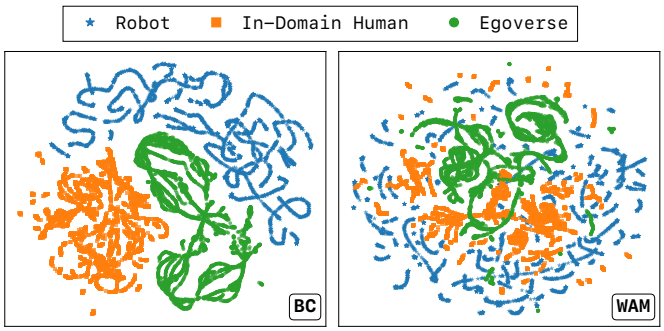


Fig. 4: **UMAP of Trunk Embeddings on Cup-on-Saucer.** BC separates human-robot (and human-human) embeddings; WAM aligns them in a shared latent.

(D1) and (D2), but remains spatially indexed on the image grid so (D3) is only partially met. (3) **Camera-stabilized 3D flow** [51, 32] predicts a dense motion field $s = F_{[t,t+T]}$ in the camera frame at t. We use a pretrained 3D point tracker with Aria VIO poses to stabilize points against head motion: $\tilde{X}_{t+T} = (T_t^{\text{cam}})^{-1} T_{t+T}^{\text{cam}} X_{t+T}$, $F = \tilde{X}_{t+T} - X_t$. Static background yields near-zero flow while manipulated objects retain physical displacement, satisfying (D1)–(D3) by construction. Head and loss details are in App. B.

Joint training, action-only inference. EGOWAM is trained end-to-end with

$$\mathcal{L}_{\text{EGOWAM}} = \mathcal{L}_{\text{action}}^{\text{robot}} + \mathcal{L}_{\text{action}}^{\text{human}} + \lambda\big(\mathcal{L}_{\text{world}}^{\text{robot}} + \mathcal{L}_{\text{world}}^{\text{human}}\big) \tag{3}$$

over mixed human/robot mini-batches. Both losses supervise the shared trunk f_ϕ: when human action labels transfer weakly, world prediction compensates and shapes the joint representation. Following [57, 28], we treat world modeling as *training-time representation shaping*: the world-model head is dropped at inference, so the policy can run at BC cost (30 Hz) while retaining the joint representation.

IV. EXPERIMENTS

We present a systematic study of EGOWAM for human-robot co-training, organized around three main questions. Full experimental setup details are in App. E. We also provide the qualitative rollout results and more experiments in App. C.

(Q1) WAM vs. BC Co-Training under Natural Human Data. Fig. 3 shows that BC co-training would harm policy performance for `cup-on-saucer` and `fold-clothes`. We trace it as robot and human action patterns are not well aligned, and thus BC degrades by introducing these *inexecutable human-like actions* (App. C, Fig. 7), whereas WAM consistently turns the same data into reliable gains. Moreover, even when large amounts of human data are introduced, BC tends to *overfit* to the robot data alone and fails to benefit from human data for generalization. App. C (Fig. 7) shows that on the fold-clothes task, BC+EgoVerse still overfits to the robot data and cannot adapt to novel scenes with lower table heights, whereas the DINO-based WAM leverages in-the-wild human data to achieve better generalization to new objects and scenes. Fig. 4 traces this to representation: BC isolates human and robot (and even human–human) embeddings under unaligned actions, whereas WAM aligns them into a shared space through additional task-relevant dynamics supervision.

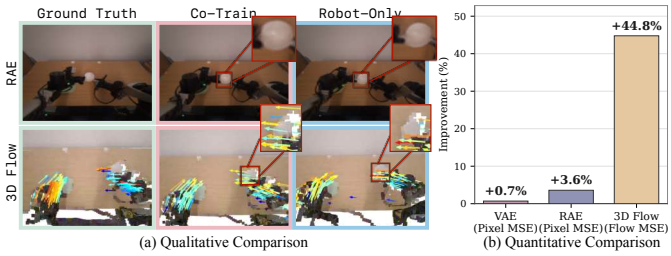


Fig. 5: **World Prediction Comparison. (a)** Co-training improves RAE and 3D Flow predictions, with better *object shape and motion*; **(b)** 3D Flow's gains far exceed Pixel and DINO from co-training.

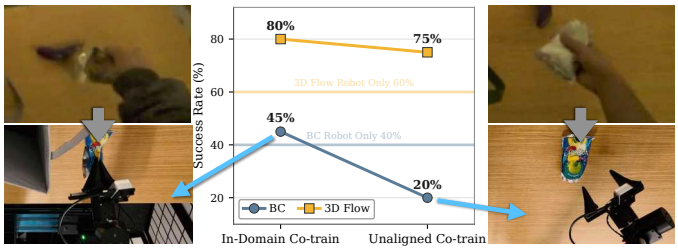


Fig. 6: **Ablation of Unaligned Human Data.** Unaligned human data collapses BC below robot-only; WAM stays robust.

(Q2) World Representation Choice for Cross-Embodiment Transfer. The world representation is the next critical axis. Fig. 3 shows that pixel-based prediction transfers weakly: reconstructing raw appearance entangles embodiment- and scene-specific detail that does not transfer, leaving the hallucination and geometric confusion in App. C (Fig. 7): a free-space grasp on bag-grocery and an unfinished final fold on fold-clothes. Abstraction fixes this in two complementary ways: the DINO-based WAM predicts semantic features and yields the strongest OOD generalization to unseen objects and scenes (Fig. 3), whereas the 3D-flow WAM predicts camera-stabilized motion and delivers the largest in-domain spatial gains, succeeding at precise cup placement across the workspace. Fig. 5 grounds this contrast: pixel (VAE) prediction barely changes, the semantic (RAE) target refines *object shape* from human data but only moderately, and the 3D-flow target benefits most, recovering *object motion* that robot-only prediction leaves static.

(Q3) Ablation: The Role of Unaligned Human Data. We further study the hypothesis that unaligned human data degrades BC. Among the three tasks, only bag-grocery benefits from BC co-training (Fig. 3), because its pick-and-place motions are naturally aligned between human and robot (Fig. 6, left). As a counterfactual, we collect unaligned human demonstrations in which the human picks objects in an unusual manner that maps to inexecutable robot actions (Fig. 6, right). Under this data, BC collapses below the robot-only baseline, whereas the 3D-flow WAM stays robust and still surpasses robot-only. The results on ablation on aligned human data is shown in App. C-A.

V. CONCLUSION

We asked what *world representation* enables transfer from egocentric human data to bimanual manipulation. EGOWAM

isolates this axis, varying only the world target. We show that WAM co-training scales more reliably than BC; pixel-VAE transfers weakly, while DINO and camera-stabilized 3D flow decisively outperform pixels with complementary strengths. EGOWAM is a starting point: the optimal world target remains open to scale egocentric human data for robot manipulation.

REFERENCES

[1] Bo Ai, Stephen Tian, Haochen Shi, Yixuan Wang, Tobias Pfaff, Cheston Tan, Henrik I Christensen, Hao Su, Jiajun Wu, and Yunzhu Li. A review of learning-based dynamics models for robotic manipulation. *Science Robotics*, 10(106):eadt1497, 2025.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

[3] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction, 2025. URL https://arxiv.org/abs/2506.21552.

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models, 2024. URL https://arxiv.org/abs/2412.03572.

[5] Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-RDT: Human Manipulation Enhanced Bimanual Robotic Manipulation. *arXiv preprint arXiv:2507.23523*, 2025.

[6] 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.

[7] 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.

[8] Daesol Cho, Youngseok Jang, Danfei Xu, and Sehoon Ha. EgoAVFlow: Robot Policy Learning with Active Vision from Human Egocentric Videos via 3D Flow. *arXiv preprint arXiv:2602.22461*, 2026.

[9] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. URL https://arxiv.org/abs/2308.13561.

[10] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi Jim Fan. DreamDojo: A Generalist Robot World Model from Large-Scale Human Videos. *arXiv preprint arXiv:2602.06949*, 2026.

[11] 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, 2026. URL https://arxiv.org/abs/2512.13644.

[12] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong

Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. *arXiv preprint arXiv:2311.18259*, 2024.

[13] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[14] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

[15] Siddhant Haldar and Lerrel Pinto. Point Policy: Unifying Observations and Actions with Key Points for Robot Manipulation. *arXiv preprint arXiv:2502.20391*, 2025.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] 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*, 2026.

[18] Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, Oier Mees, Marc Pollefeys, Zhuang Liu, Jiajun Wu, Pieter Abbeel, Jitendra Malik, Yilun Du, and Jianfei Yang. World model for robot learning: A comprehensive survey. 2026.

[19] Adam Hung, Bardienus Pieter Duisterhof, and Jeffrey Ichnowski. 3PoinTr: 3D Point Tracks for Robot Manipulation Pretraining from Casual Videos. *arXiv preprint arXiv:2603.08485*, 2026.

[20] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[21] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling Imitation Learning via Egocentric Video. *arXiv preprint arXiv:2410.24221*, 2024.

[22] 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.

[23] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning. *arXiv preprint arXiv:2601.16163*, 2026.

[24] Seungjae Lee, Yoonkyo Jung, Inkook Chun, Yao-Chih Lee, Zikui Cai, Hongjia Huang, Aayush Talreja, Tan Dat Dao, Yongyuan Liang, Jia-Bin Huang, and Furong Huang. TraceGen: World Modeling in 3D Trace Space Enables Learning from Cross-Embodiment Videos. *arXiv preprint arXiv:2511.21690*, 2025.

[25] 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*, 2025.

[26] Hongyu Li, Lingfeng Sun, Yafei Hu, Duy Ta, Jennifer Barry, George Konidaris, and Jiahui Fu. NovaFlow: Zero-Shot Manipulation via Actionable Flow from Generated Videos. *arXiv preprint arXiv:2510.08568*, 2025.

[27] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-Life Human Activity Videos. *arXiv preprint arXiv:2510.21571*, 2025.

[28] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified Video Action Model. *arXiv preprint arXiv:2503.00200*, 2025.

[29] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[30] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[31] 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.

[32] Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4world: Feedforward world-centric dense 3d tracking of all pixels. *arXiv preprint arXiv:2603.02573*, 2026.

[33] Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. DiT4DiT: Jointly Modeling Video Dynamics and Actions for Generalizable Robot Control. *arXiv preprint arXiv:2603.10448*, 2026.

[34] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured World Models from Human Videos. *arXiv preprint arXiv:2308.10901*, 2023.

[35] 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.

[36] Nilaksh, Saurav Jha, Artem Zholus, and Sarath Chandar. Reconstruction or semantics? what makes a latent space useful for robotic world models. 2026. URL https://arxiv.org/abs/2605.06388.

[37] Dantong Niu, Yuvan Sharma, Haoru Xue, Giscard Biamby, Junyi Zhang, Ziteng Ji, Trevor Darrell, and Roei Herzig. Pre-training auto-regressive robotic models with 4d representations. *arXiv preprint arXiv:2502.13142*, 2025.

[38] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[39] Jedrzej Orbik and Frederik Ebert. Oculus reader: Robotic teleoperation interface. https://github.com/rail-berkeley/oculus_reader, 2021.

[40] Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-Action Models for Generalizable Robot Control Beyond VLAs. *arXiv preprint arXiv:2512.15692*, 2025.

[41] 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*, 2025.

[42] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Chenyu Yang, Xi Wang, James Fort, Richard Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. EgoVerse: An Egocentric Human Dataset for Robot Learning from Around the World. *arXiv preprint arXiv:2604.07607*, 2026.

[43] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy ˜ human policy. *arXiv preprint arXiv:2503.13441*, 2025.

[44] Juntao Ren, Priya Sundaresan, Dorsa Sadigh, Sanjiban Choudhury, and Jeannette Bohg. Motion Tracks: A Unified Representation for Human-Robot Transfer in Few-Shot Imitation Learning. *arXiv preprint arXiv:2501.06994*, 2025.

[45] Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

[46] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[47] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers. *arXiv preprint arXiv:2409.20537*, 2024.

[48] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025.

[49] Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive World Simulator for Robot Policy Training and Evaluation. *arXiv preprint arXiv:2603.08546*, 2026.

[50] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point Trajectory Modeling for Policy Learning. *arXiv preprint arXiv:2401.00025*, 2024.

[51] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D Point Tracking Made Easy. *arXiv preprint arXiv:2507.12462*, 2025.

[52] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the Cross-Domain Manipulation Interface. *arXiv preprint arXiv:2407.15208*, 2024.

[53] 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.

[54] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen

Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World Action Models are Zero-shot Policies. *arXiv preprint arXiv:2602.15922*, 2026.

[55] Zhao-Heng Yin, Sherry Yang, and Pieter Abbeel. Object-centric 3D Motion Field for Robot Learning from Human Videos. *arXiv preprint arXiv:2506.04227*, 2025.

[56] 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.

[57] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do World Action Models Need Test-time Future Imagination? *arXiv preprint arXiv:2603.16666*, 2026.

[58] Kevin Zakka. Mink: Python inverse kinematics based on MuJoCo. https://github.com/kevinzakka/mink, 2025.

[59] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[60] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Particle-grid neural dynamics for learning deformable object models from rgb-d videos. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[61] Kaifeng Zhang, Shuo Sha, Hanxiao Jiang, Matthew Loper, Hyunjong Song, Guangyan Cai, Zhuo Xu, Xiaochen Hu, Changxi Zheng, and Yunzhu Li. Real-to-sim robot policy evaluation with gaussian splatting simulation of soft-body interactions. *arXiv preprint arXiv:2511.04665*, 2025.

[62] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[63] Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling Dexterous Manipulation with Diverse Egocentric Human Data. *arXiv preprint arXiv:2602.16710*, 2026.

[64] Hongyan Zhi, Peihao Chen, Siyuan Zhou, Yubo Dong, Quanxi Wu, Lei Han, and Mingkui Tan. 3DFlowAction: Learning Cross-Embodiment Manipulation from 3D Flow World Model. *arXiv preprint arXiv:2506.06199*, 2025.

[65] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. *arXiv preprint arXiv:2411.04983*, 2025.

[66] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets. *arXiv preprint arXiv:2504.02792*, 2025.

[67] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A Modular Simulation Framework and Benchmark for Robot Learning. *arXiv preprint arXiv:2009.12293*, 2020.

Supplementary Materials

Contents

A Related Work **9**
B Method Details **9**
B-A Aligned Action Channel: Additional Detail . . . 9
B-B WAM as a World-Level Transfer Interface 10
B-C Single-Backbone Architecture and Controlled Variables 10
B-D World-Model Target Instantiations 10
B-E Joint Training and Action-Only Inference 10
C Additional Real-World Experiment Results and Analysis **10**
C-A Ablation on Aligned Human Data 10
C-B Ablation on Human Data Modality 11
C-C Failure Analysis for Pixel-PT on Bag-Grocery 12
D Additional Implementation Details **12**
D-A Behavior-Cloning Co-Training Architecture . . . 12
D-B World-Model Heads 12
D-C Training and Inference 14
E Additional Real-World Experiment Details **15**
E-A Robot Platform and Data Collection 15
E-B Human Data Collection and EgoVerse Dataset 15
E-C Rollout Evaluation Protocol 15
F Simulation Experiments **16**
F-A Benchmark and Task Setup 16
F-B Integration into EgoWAM 16
F-C World-Model Variants and Training 16
F-D Closed-Loop Evaluation Protocol 17
F-E Results 17

A. Related Work

Robot Learning from Human Data. Human video is a scalable data source for robot manipulation [12, 17, 42]. One line of work treats it as a pretraining corpus for action-aware visual representations later transferred to downstream planning or policy [35, 2, 37, 3]. Beyond pretraining, prior work uses human data more directly for manipulation along two axes. The first retargets human demonstrations into a shared action space and co-trains an imitation policy [21, 43, 53, 5, 56, 27] or world model [11, 18, 34] on both human and robot data; because of domain gaps across embodiments, this hinges on tight alignment in viewpoint, speed, and kinematics [63, 25, 31, 41, 6]. The second axis sidesteps the action decoder by extracting object or scene motion from human video and decoding actions from it at inference, especially with 2D point trajectory tracks [50, 52, 44, 15, 24] and 3D flow motion fields [19, 64, 26, 55, 8]. EgoWAM departs from both: scene motion supervises the shared trunk during *end-to-end* training and is discarded at inference, transferring task-relevant dynamics without injecting inexecutable motions or adding test-time cost.

World Action Models for Robot Learning. A world model predicts how the environment evolves given an action [13, 14, 1, 4]; a World Action Model (WAM) couples such a predictor to a policy, grounding action prediction in task-relevant dynamics [18, 28, 66]. Most existing WAMs build on a pretrained video model to offer a spatio-temporal representation for action decoding [54, 23, 40, 33, 7], while a parallel line instead treats the video/world model as an interactive simulator for data generation or policy evaluation [10, 49, 61]. A central but unresolved question for generalizable WAMs is *what world representation* supports scaling from diverse data [65, 36, 59, 60]. We study this question directly inside the WAM co-training framework, comparing three representations under a single training-as-representation-shaping setup [57, 28]: reconstruction-based pixel latents [46], semantics-based DINO features [38, 62, 45], and camera-stabilized 3D motion flow [51, 32].

B. Method Details

Sec. II–III summarize the EgoWAM problem setup, aligned action channel, WAM interface, desiderata, backbone design, world-target choices, and training objective (Eqs. 1–3). This appendix supplies the derivations, architectural specifics, per-head loss definitions, and training protocol omitted from the main text for space. Implementation hyperparameters and head dimensions are in App. D.

A. Aligned Action Channel: Additional Detail

We align the action channel as much as possible so BC co-training is a strong baseline [42]. Robot actions $a^R_{t:t+k} \in \mathbb{R}^{k\times d_a}$ are computed from joint angles via forward kinematics and re-expressed in the static ego-camera frame. Human hand poses $p^H_{t+i} \in \mathrm{SE}(3)$ are natively expressed in the moving device frame T^{device}_t; we re-express them in the *instantaneous* device frame at time t,

$$a^H_{t:t+k} = \left[(T^{\text{device}}_t)^{-1} T^{\text{device}}_{t+i}\, p^H_{t+i}\right]_{i=1}^{k} \tag{4}$$

factoring out head ego-motion so the same physical motion yields a comparable numerical trajectory across embodiments.

Two residual mismatches remain after coordinate alignment. *(i) Speed.* Humans act faster than teleoperated robots, so we use embodiment-specific windows spanning comparable progress—$T_H = 1\,\text{s}$, $T_R = 1.5\,\text{s}$—both discretized into k steps. We use k for the resampled chunk length and $T \in \{T_H\ T_R\}$ for the original-time horizon; world-model targets s_{t+T} are indexed in original time. *(ii) Workspace range.* Quantile normalization maps each action dimension's 1st and 99th percentiles to $[-1\ 1]$, robust to hand-tracking outliers.

Eq. 1 is realized as the sum of per-embodiment conditional flow-matching losses, $\mathcal{L}_{\text{BC-cotrain}} = \mathcal{L}^{\text{robot}}_{\text{CFM}} + \mathcal{L}^{\text{human}}_{\text{CFM}}$. This is the strongest action-aligned baseline our system supports; the remaining transfer gap motivates the world-model interface below.

B. WAM as a World-Level Transfer Interface

Eq. 2 factorizes as a joint model over action and future-state read-outs of a shared latent $z_t = f_\phi(o_t)$:

$$p_{\theta,\psi}\big(a_{t:t+k} \;\; s_{t+T} \,\big|\, o_t\big) = p_\psi\big(s_{t+T} \mid z_t\big)\, p_\theta\big(a_{t:t+k} \mid z_t\big) \quad (5)$$

where λ in Eq. 2 trades off action fidelity against future-prediction supervision. The key property is that $\mathcal{L}_{\text{world}}$ supervises the *shared trunk* through dynamics produced by *both* embodiments, while $\mathcal{L}_{\text{action}}$ supervises through the aligned actions of Sec. B-A. Human data can therefore shape the shared representation z_t through future-scene prediction even when its action labels do not transfer faithfully. The world target s_{t+T} decides whether human and robot data converge to a shared representation or are pulled apart by embodiment-specific signal.

C. Single-Backbone Architecture and Controlled Variables

EGOWAM builds on the Heterogeneous Pretrained Transformer (HPT) backbone [47, 42], where embodiment-specific stems tokenize each input into a shared latent space and a single transformer trunk operates on the unified token stream. Observation, action, and future-prediction tokens coexist in one stream and attend within the same trunk, exposing future dynamics and action prediction as two read-outs of a shared latent. We inherit the tokenizer and trunk and add a second bank of learnable *future tokens* and a *swappable* world-model head that consumes them (Fig. 2).

Heterogeneous tokenizers. Each embodiment is tokenized through a shallow, embodiment-specific stem. A shared ego-vision stem encodes I_t^{ego} with a ResNet-18 encoder [16]; a matching wrist-vision stem encodes I_t^{wrist} on robot batches. The proprioception stem encodes the end-effector pose with a per-embodiment MLP. Shared learnable *action* and *future* tokens query the trunk, which routes them to the action and world-model heads.

Action head. The action head π_θ is a multi-block transformer decoder trained with conditional flow matching [29, 42]. Given a clean chunk $a_{t:t+k}$ and noise $\epsilon \sim \mathcal{N}(0\ I)$, we draw $\tau \sim \text{Beta}(1\ 5\ 1\ 0)$ and form $a^\tau_{t:t+k} = (1-\tau)\epsilon + \tau\, a_{t:t+k}$ to initialize the action tokens, with τ embedded along the hidden dimension. Alternating self- and cross-attention blocks denoise the tokens while injecting trunk context, and a linear layer projects them into the action dimension.

Swappable world-model head. The world-model head g_ψ is conditioned on the trunk embeddings and predicts the target s_{t+T} at the embodiment-specific horizon T (Sec. B-A). We vary the world representation while holding the trunk, action head, and data mixture fixed.

D. World-Model Target Instantiations

Sec. III introduces the three targets and their relation to desiderata (D1)–(D3). Here we specify the per-head objectives; all three share the linear flow path $s^\tau = (1-\tau)\epsilon + \tau s$, $\tau \in [0\ 1]$, $\epsilon \sim \mathcal{N}(0\ I)$.

Pixel VAE (reconstruction baseline). The Pixel-VAE variant predicts a future ego frame in the latent space of a pretrained video VAE [46]: $s = \text{VAE}(I^{\text{ego}}_{t+T})$. The head is a diffusion transformer (DiT) following VACE-1.3B [20] without text conditioning, initialized either from pretrained weights (**Pixel-PT**) or from scratch (**Pixel**), and trained to predict noise: $\mathcal{L}^{\text{VAE}}_{\text{world}} = \mathbb{E}\big\|\epsilon - \epsilon_\psi(s^\tau\ \tau\ f_\phi(o))\big\|^2$.

DINO features (semantic abstraction). The target is DINO [38] patch features of the ego frame at $t{+}T$: $s = \text{DINO}(I^{\text{ego}}_{t+T})$. The head follows the wide-DDT design of RAE [62], which widens the denoiser to match the high channel count of semantic latents and decouples noise-prediction from feature-conditioning blocks: $\mathcal{L}^{\text{RAE}}_{\text{world}} = \mathbb{E}\big\|\epsilon - \epsilon_\psi(s^\tau\ \tau\ f_\phi(o))\big\|^2$. DINO features remain spatially indexed in image coordinates, so head ego-motion (D3) is only partially mitigated.

3D flow (spatial abstraction). The target is a dense 3D motion field over $[t\ \ t+T]$ in the camera-stabilized frame at time t: $s = F_{[t,t+T]}$. We feed the pretrained 3D point tracker [51, 32] with Aria VIO camera poses [9], map the future position back to the camera frame at t, $\tilde{X}_{t+T} = (T^{\text{cam}}_t)^{-1}\, T^{\text{cam}}_{t+T}\, X_{t+T}$, and define $s = F_{[t,t+T]} = \tilde{X}_{t+T} - X_t$. After stabilization, static background yields near-zero flow while manipulated objects retain motion proportional to physical displacement. The head is an alternating self/cross-attention decoder that takes query points q uniformly sampled from the current ego frame and regresses velocity: $\mathcal{L}^{\text{Flow}}_{\text{world}} = \mathbb{E}\big\|u_\psi(s^\tau\ \tau\ f_\phi(o)\ \ q) - (s_q - \epsilon_q)\big\|^2$. Head dimensions, grid resolution, and movement thresholds are in App. D.

E. Joint Training and Action-Only Inference

Eq. 3 is optimized over mixed human/robot mini-batches. Each step draws samples from $\mathcal{D}_R$ and $\mathcal{D}_H$; both pass through the shared tokenizers and trunk, after which the action head yields $\mathcal{L}_{\text{action}}$ (Eq. 1) and the world-model head yields $\mathcal{L}_{\text{world}}$. Both losses supervise the shared trunk f_ϕ: when human action labels transfer weakly, world prediction compensates and shapes the trunk representation. At inference we drop the world-model head and unroll only the action head to decode $a_{t:t+k}$ from the trunk embeddings, matching the latency of a same-size BC policy (30 Hz) while retaining the cross-embodiment representation learned under joint supervision [57, 28].

C. Additional Real-World Experiment Results and Analysis

A. Ablation on Aligned Human Data

Sec. IV**(Q3)** studied one extreme of action alignment: *deliberately misaligned* human data collapses BC below its robot-only baseline, while the 3D-flow WAM stays robust. Here we probe the other extreme: does *manually aligning* the demonstrator to the robot's viewpoint and grasp strategy (Fig. 9(c)) help each paradigm? We co-train on `cup-on-saucer`, the task with the strongest human–robot mismatch (Fig. 1), under two 1:1 human regimes: **In-Domain Co-train** (natural human data, unmatched viewpoint and behavior studied in Sec. IV) and **Aligned Co-train** (human demonstrator intentionally mirrors the robot, described in App. E).

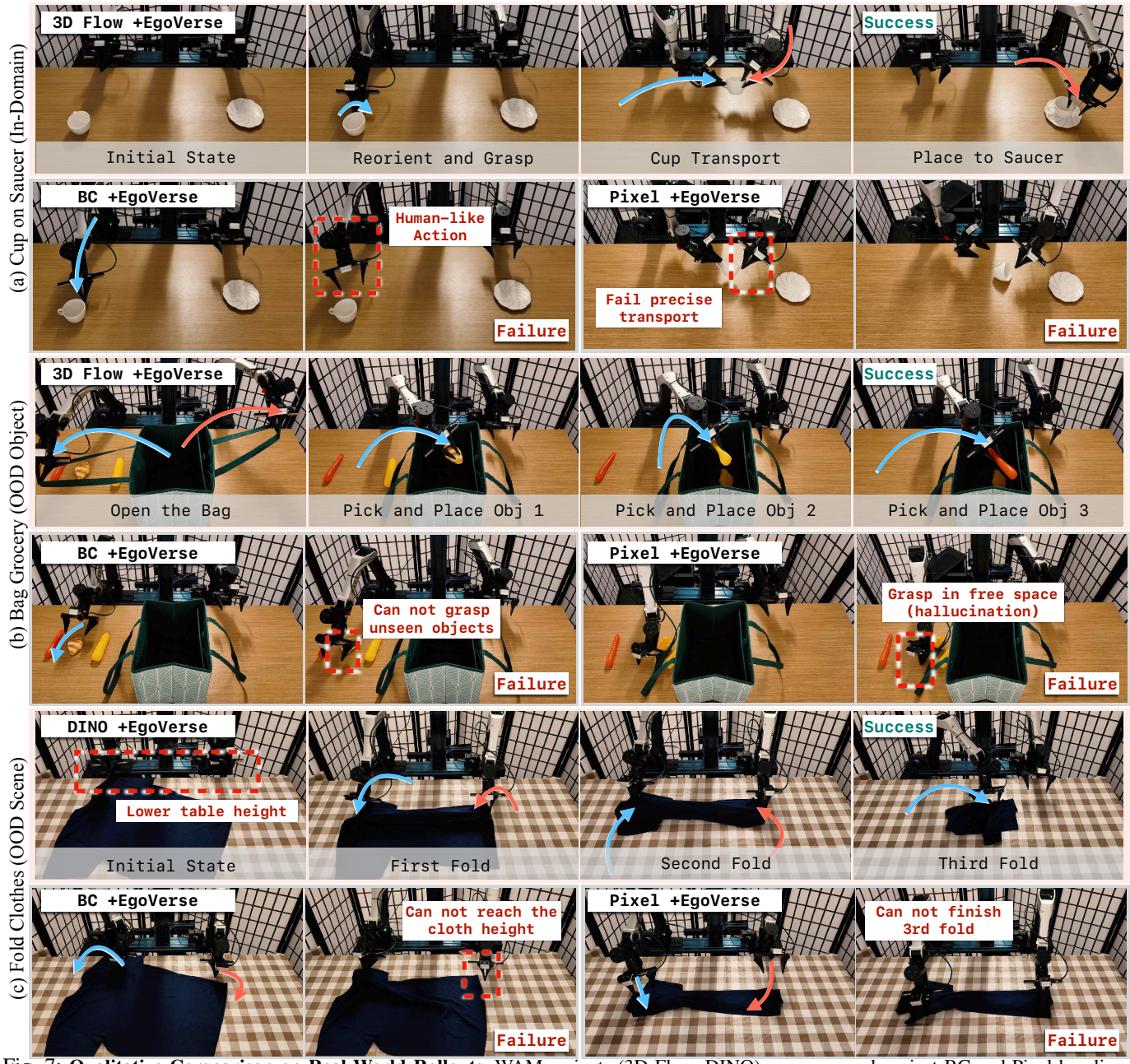


Fig. 7: **Qualitative Comparison on Real-World Rollouts.** WAM variants (3D Flow, DINO) are compared against BC and Pixel baselines under *EgoVerse co-training*. **(a)** BC produces human-like motion, Pixel fails at precise cup transport; **(b)** BC cannot grasp unseen objects, Pixel hallucinates a free-space grasp; **(c)** BC overfits and fails to reach the cloth, Pixel's geometric confusion leaves the third fold incomplete.

Fig. 8 reports the difference between In-Domain Co-train and Aligned Co-train across the four variants (BC, Pixel, DINO, 3D Flow). Three findings stand out:

- **Aligned human data lifts BC above its robot-only baseline.** BC drops under natural in-domain human data, reproducing the negative-transfer pattern. When the demonstrator is aligned to the robot, BC effectively gains from human data, confirming that BC can benefit from human data *only* when the action distribution is hand-curated to match the robot's.
- **Pixel and DINO improve when ego-motion is factored out at collection time.** Pixel rises $35\% \rightarrow 65\%$ and DINO $50\% \rightarrow 70\%$ when the human head motion matches the robot's static ego-camera. The gap quantifies how much these image-coordinate targets suffer when *D3 (ego-motion factoring)* is violated.
- **The 3D-flow WAM is invariant to alignment.** It holds at 85% under both regimes: the camera-stabilized 3D-flow target factors out ego-motion by construction (Sec. B-D), so the gains others recover through manual alignment come for free.

B. Ablation on Human Data Modality

How much of EgoWAM's gain comes from the *action labels* on human demonstrations versus the *world-model supervision*? We ablate three human co-training modalities on

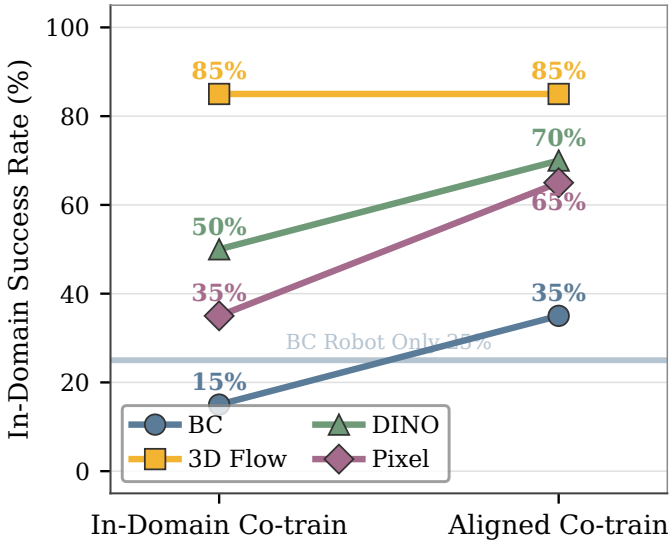


Fig. 8: **Ablation of Aligned Human Data.** Aligning the demonstrator lifts BC above robot-only and gains Pixel/DINO 20–30 pts, isolating human head motion; 3D-flow holds at 85% either way.

`cup-on-saucer`: **Action Only** (BC co-train baseline), **3D Flow only** (no action supervision on human batches), and **Action + 3D Flow** (full EGOWAM). Fig. 10 yields two findings:

- **World-model supervision alone outperforms action supervision alone.** 3D-flow-only outperforms action-only across all three splits, most starkly on OOD Scene, where action-only fails entirely (0%) while 3D-flow-only retains 10% SR. The trunk shaping induced by predicting 3D motion in human video is a stronger context-level transfer signal than action labels.
- **Action labels and world-model supervision are mutually reinforcing.** Joint training wins on every split, and we attribute this to two complementary effects. (1) *Action as context*: action labels condition the trunk on the demonstrator's intent, sharpening the world model prediction; this effect is measurable directly in the world-model loss, where co-training lowers the 3D-flow prediction loss on both the human and robot streams (Table I). (2) *Action as task-relevance signal*: action labels mark which motion in the scene is task-relevant, focusing the trunk on agent-caused dynamics rather than incidental flow.

TABLE I: 3D-flow world-model prediction loss. Adding action labels (*Action + Flow*, full EGOWAM) lowers the loss over *Flow-Only* on both streams, supporting *action as context*.

Flow stream	Flow-Only	Action + Flow
Human	0.23	**0.22**
Robot	0.20	**0.19**

C. Failure Analysis for Pixel-PT on Bag-Grocery

As shown in Fig. 3, Pixel-PT (robot only) underperforms both BC and Pixel on `bag-grocery` task. Here, we provide further failure analysis on it. From our observation, most failures occurred at the bag-opening stage, and Fig. 11 isolates the cause. **Pixel-PT** renders crisp future frames in which the bag already appears open before the gripper has acted on the handles: the natural-image prior, in which bags are typically open, overrides the actual scene state, and the policy advances to the pick stage prematurely. **Pixel (from-scratch)** produces blurrier frames but tracks bag openness faithfully, and the resulting policy opens the bag more reliably. **Pixel-PT + EgoVerse** keeps the sharp pretrained output while predicting the closed-then-opening progression correctly, indicating that human co-training can correct noisy prior-induced hallucination in pretrained video models by increasing the data amount, and thus improves the policy a lot.

D. Additional Implementation Details

A. Behavior-Cloning Co-Training Architecture

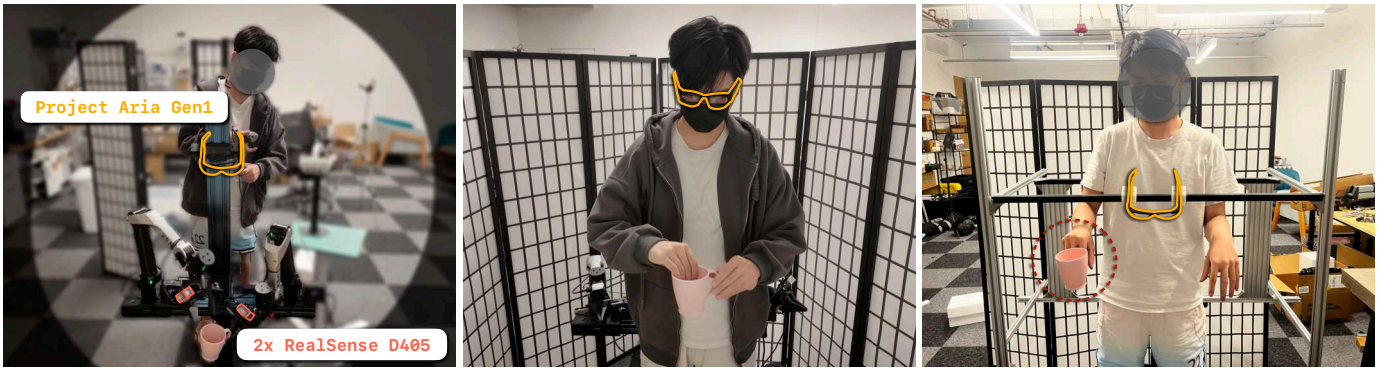
Stems and Trunk. Each embodiment is tokenized by a shallow, embodiment-specific stem that projects into a shared 256-d space via learned query attention (16 latent queries, 8 heads, head dim 64). The egocentric stream uses a ResNet-18 [16] stem shared across embodiments; robot batches add a matching wrist-view stem, and the 14-d end-effector proprioception is encoded by a per-embodiment MLP. A single transformer trunk (256-d, 16 blocks, 8 heads, stochastic depth 0 1, with learned domain embeddings) processes the observation tokens together with 64 learnable action tokens and 16 future tokens, using a one-frame observation history.

Flow-Matching Action Head. The action head is a 6-block CrossTransformer (hidden width 128, 4 heads) trained with the conditional flow-matching objective of Sec. B-C. It denoises a chunk of 100 action tokens via alternating self- and cross-attention conditioned on the trunk, with the flow-matching timestep drawn from $\mathrm{Beta}(1\ 5\ \ 1\ 0)$, and is sampled with 50 v-prediction steps. Cross-embodiment actions are unified to the 14-d end-effector space and quantile-normalized (Sec. B-A).

B. World-Model Heads

All three heads share the linear flow path $s^\tau = (1-\tau)\epsilon + \tau s$, $\tau \in [0\ \ 1]$, $\epsilon \sim \mathcal{N}(0\ \ I)$, and differ only in the target s_{t+T} and the head architecture that consumes the trunk embeddings.

VAE Video Prediction Head. The target is the future ego frame (resized to 128×128) encoded in the latent space of a frozen pretrained Wan video VAE [46], $s = \mathrm{VAE}(I^{\mathrm{ego}}_{t+T})$, giving a 16×16×16 latent. We study two instantiations of this head, both trained under flow-matching v-prediction with 50 sampling steps. **Pixel** predicts the latent with a lightweight diffusion transformer (DiT) of 6 blocks, hidden width 384, and 6 heads, patchified at stride 2 (8×8 = 64 tokens) and conditioned on the mean-pooled trunk embedding, trained from scratch. **Pixel-PT** is initialized with the pretrained Wan 1.3B transformer. This backbone has 30 layers, hidden width 1536, 12 attention heads, and FFN dimension 8960. The two variants share the same VAE target and trunk conditioning, and differ in head capacity and initialization.



(a) Robot Teleoperation (b) Natural Human Demonstration (c) Robot-Aligned Human Demonstration

Fig. 9: **Robot Platform and Data Collection. (a)** Our robot platform: two upright-mounted 6-DoF ARX5 arms with two wrist-mounted Intel RealSense D405 cameras and a head-mounted Project Aria headset, teleoperated by a human via a Meta Quest 3 interface. **(b)** A human wearing Project Aria glasses collects demonstrations naturally. **(c)** The human deliberately aligns with the robot's action, height, and viewpoint, producing a static egocentric view matched to the robot platform.

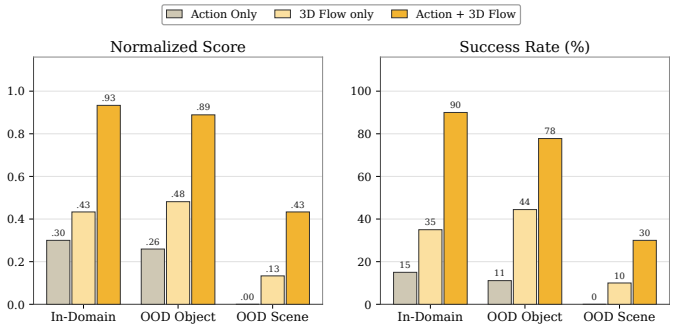


Fig. 10: **Ablation of Co-Train Human Data Modality.**

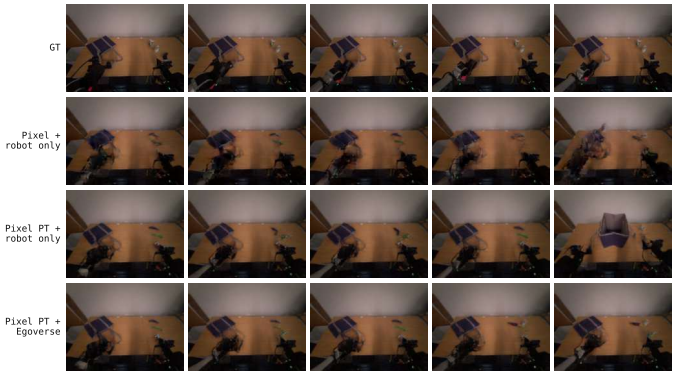


Fig. 11: **Video Prediction for Failure Analysis on Bag-Grocery.** 6-step rollouts from the same initial frame. **Pixel + robot only**: blurry but faithful, where the bag stays closed until acted on. **Pixel-PT + robot only**: hallucinates an already-open bag, causing the policy to skip the opening stage. **Pixel-PT + Egoverse**: sharp and faithful, showing human co-training removes the hallucination.

RAE DINO Prediction Head. The target is the DINOv2-B [38] patch-feature map of the future ego frame, $s = \mathrm{DINO}(I^{\text{ego}}_{t+T})$, a 16×16 token grid in $\mathbb{R}^{768}$. We adopt the Representation Autoencoder (RAE) [62] on two fronts: its frozen DINOv2 encoder defines the prediction target, and its DiT^{DH} denoiser architecture defines our head. The trained pixel decoder is not used during EGOWAM training – supervision happens directly in feature space and the world-model head is discarded at inference – and is only invoked for visualization. Following RAE, we drop `[CLS]` and register tokens and apply per-token layer normalization before the loss. The DiT^{DH} head pairs a standard DiT backbone with a *shallow but wide* DDT-style head [48], motivated by RAE's finding that diffusion in semantic latent spaces requires denoiser width $\geq$ token dimensionality (768 here), below which the flow-matching loss provably fails to converge. We use a 6-block, 384-d backbone (6 heads) and a 2-block, 2048-d wide head, both conditioned on the mean-pooled trunk embedding, and train under v-prediction: $\mathcal{L}^{\text{DINO}}_{\text{world}} = \mathbb{E}\,\|\epsilon - \epsilon_\psi(s^\tau \;\; \tau \;\; f_\phi(o))\|^2$, sampled with 50 fixed Euler steps.

3D Flow Data Processing and Prediction Head. The target is a dense 3D motion field over $[t \;\; t+T]$ expressed in the camera-stabilized frame at time t. We obtain it with Track4World [32], a feed-forward dense 3D point tracker that takes RGB frames alone and predicts per-pixel 3D scene flow together with metric depth, intrinsics, and camera poses. Raw 3D displacement on egocentric video is dominated by ego-motion: stationary objects induce large apparent flow whenever the wearer turns their head. We therefore use the Aria VIO head poses [9] to re-express the predicted future positions X_{t+T} in the camera frame at t, $\tilde{X}_{t+T} = (T^{\text{cam}}_t)^{-1}\, T^{\text{cam}}_{t+T}\, X_{t+T}$, and define the flow target $s = F_{[t,t+T]} = \tilde{X}_{t+T} - X_t$. After this stabilization a static background yields near-zero flow while manipulated objects retain motion proportional to physical displacement; for robot clips the head camera is fixed, so the transform is the identity. The flow is read on a fixed 28×40 (1120-point) pixel-anchor grid, and to suppress tracking noise we discard anchors whose displacement falls below a movement threshold (2 mm for robot tracks; 10 mm for human tracks, which carry residual head motion) and skip the first/last 20 frames of each human clip. The head is a flow-matching decoder (4 blocks, hidden width 256, 4 heads) that, conditioned on the anchor positions q and the trunk features, regresses the 3D displacement of *all* 1120 anchors over the 100-step horizon (target shape 100×1120×3) with no subsampling, $\mathcal{L}^{\text{Flow}}_{\text{world}} = \mathbb{E}\,\|u_\psi(s^\tau \;\; \tau \;\; f_\phi(o) \;\; q) - (s_q - \epsilon_q)\|^2$.

TABLE II: Hyperparameters held fixed across all world-model variants. Only the world-model head and target (bottom block) change between runs.

Component / Setting	**Value**
Backbone and stems	
Trunk (embed dim / blocks / heads)	$256/16/8$
Query tokens consumed by trunk	64 action + 16 future (WAM); 64 action only (BC)
Ego / wrist vision stem	ResNet-18 [16], output dim 256
Proprioception stem	MLP ($14 \to 256$)
Cross-attn stem (latent / heads / head dim)	$16/8/64$
Observation horizon	1
Stochastic depth (drop path)	0.1
Action head	
Architecture	CrossTransformer (flow matching)
Blocks / hidden dim / heads	$6/128/4$
Action dim d_a	14 (per-arm 6-DoF SE(3) + gripper)
Resampled chunk length k	100
Robot / human window T_R, T_H	1.5 s (45 frames) / 1.0 s (30 frames)
Prediction type / sampling steps	v-prediction / 50
Flow-matching τ prior	$\mathrm{Beta}(1.5, 1.0)$
Normalization	quantile (1st/99th pct. $\to [-1, 1]$)
Deployment inference rate	30 Hz
Optimization	
Optimizer	AdamW (lr 1×10^{-4}, weight decay 1×10^{-4})
Scheduler	cosine annealing ($T_{\max} = 1400$, $\eta_{\min} = 1\times10^{-5}$)
Batch size (robot / human)	$32/32$
Steps per epoch / max epochs	$100/2000$
Precision	bf16
World-model loss weight λ	1.0
Augmentation (train only)	color jitter $(0.1, 0.1, 0.1, 0.05)$
Image preprocessing	ImageNet normalization
World-model heads (swappable)	
Pixel: VAE / decoder	Wan VAE [46] / DiT, from scratch
decoder (blocks / dim / heads)	$6/384/6$
latent ($C\times H\times W$) / input res	$16\times16\times16$ / 128
Pixel-PT: VAE / decoder	Wan VAE [46] / VACE-1.3B [20], pretrained
decoder (layers / dim / heads / FFN)	$30/1536/12/8960$
latent ($C\times H\times W$) / input res	$16\times16\times16$ / 128
DINO: encoder / pre-processing	DINOv2-B [38], frozen / drop `[CLS]` & registers, per-token LN
denoiser ($\mathrm{DiT}^{\mathrm{DH}}$ [62])	DiT backbone + wide DDT-style head [48]
backbone (blocks / dim / heads)	$6/384/6$
wide head (blocks / dim)	$2/2048$
feature dim ($C\times H\times W$)	$768\times16\times16$
3D Flow: tracker / denoiser	Track4World [32] / flow matching
denoiser (blocks / dim / heads)	$4/256/4$
anchor grid / dim / horizon	28×40 (1120 pts, no subsampling) / 3 / 100
query condition	anchor positions q from current ego frame

C. Training and Inference

Training. Each step draws 32 robot and 32 human samples and supervises the shared trunk with $\mathcal{L}_{\text{action}} + \lambda\,\mathcal{L}_{\text{world}}$, $\lambda = 1$ (Eq. 3). We optimize with AdamW (learning rate 1×10^{-4}, weight decay 1×10^{-4}) under cosine annealing ($T_{\max} = 1400$, $\eta_{\min} = 1\times10^{-5}$) in bf16. All variants except Pixel-PT are trained on a **single NVIDIA L40S GPU** for 2000 epochs of 100 steps; Pixel-PT is trained on $2\times$ L40S in data-parallel for 1000 epochs of 100 steps to accommodate the 1 3B-parameter pretrained backbone within memory. Under either configuration, end-to-end training takes approximately two days per task per method. Images are ImageNet-normalized at train and test time, with color jitter (±0 1 brightness/contrast/saturation, ±0 05 hue) added during training only. Remaining settings are listed in Table II.

Inference. At deployment the world-model head g_ψ is frozen and detached from the computation graph: only the shared trunk f_ϕ and the flow-matching action head π_θ are unrolled to decode $a_{t:t+k}$ from a one-frame observation, consistent with the action-only formulation in Sec. B-E. The full policy runs on a **single NVIDIA RTX 4090 GPU at** 30 **Hz**, matching the control rate of the underlying ARX5 platform. Discarding the world-model head at inference also keeps the deployment footprint identical across all four world-representation vari-

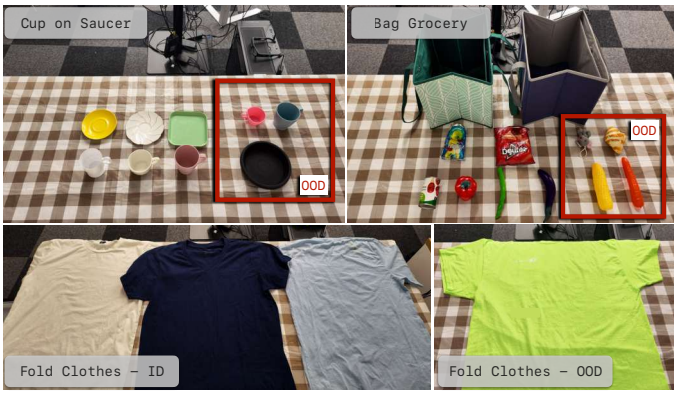


Fig. 12: **In-Domain and OOD Objects.**

ants, so any rollout differences in Sec. IV reflect what each target taught the trunk during training rather than test-time compute.

E. Additional Real-World Experiment Details

A. Robot Platform and Data Collection

As shown in Fig. 9, our bimanual platform comprises two upright-mounted 6-DoF ARX5 arms with parallel-jaw grippers, head-mounted Project Aria Gen-1 glasses [9] providing the egocentric RGB stream shared with human demonstrations, and two wrist-mounted Intel RealSense D405 cameras.

We collect demonstrations through the RAIL Lab Oculus Reader [39] interface, driving the arms with a Meta Quest 3 headset and Touch Pro controllers. Commanded base-frame end-effector poses are converted to joint angles by the Mink IK solver [58], and the resulting targets are tracked by the ARX5 joint-space controller. The runtime is a multi-threaded Python stack; low-level hardware communication runs over a CAN bus.

Robot actions are per-arm 6-DoF end-effector poses with a 1-D gripper command, $a^R_{t:t+k} \in \mathbb{R}^{k\times 14}$. They are computed from commanded joint angles via forward kinematics, projected into the egocentric Aria camera frame using the calibrated extrinsics, and expressed as $(x\ y\ z$ yaw pitch roll$)$ Euler poses per arm. Following Sec. B-A, actions are quantile-normalized by mapping each dimension's 1st and 99th percentiles to $[-1\ 1]$ to be robust to tracking outliers.

We collect 300–360 teleoperated demonstrations per task, with randomized object placements, orientations, and 4–8 object combinations for each task. Per-task demonstration counts and hours are listed in Table III, and the training/testing object sets are shown in Fig. 12.

B. Human Data Collection and EgoVerse Dataset

In-Domain Human Data Collection Setup. In-domain human data is captured with Project Aria glasses, lightweight (75 g) head-worn devices with a wide-FoV RGB camera and two synchronized monochrome scene cameras used for SLAM. Hand poses (21 keypoints and 6-DoF palm pose per hand) and a calibrated 6-DoF head pose from visual–inertial SLAM are recovered through the Aria Machine Perception Service (MPS) [9]. Following the EgoVerse protocol, demonstrations are recorded in roughly 5-minute units yielding several demonstrations each, within an approximately $40\,\text{cm} \times 60\,\text{cm}$ workspace with object positions randomized across trials. In-domain human data uses the *same* scenes and objects as the robot data but differs in viewpoint and behavior, and is collected at a 1:1 ratio with the robot demonstrations (Table III).

TABLE III: Data composition per task. *In-Domain* denotes small-scale, scene- and object-aligned human data; *EgoVerse* denotes the large-scale EgoVerse-A flagship split.

Task	Robot (# demos / h)	Human (In-Dom.)	Human (EgoV.)
cup-on-saucer	300 / 2.5h	2h	20.5h
fold-clothes	360 / 3.0h	2h	21h
bag-grocery	300 / 2.5h	2h	7h

Aligned and Unaligned Human Data Collection. To probe the robustness of each paradigm to action-(mis)aligned human data (Sec. IV**(Q3)** and App. C-A), we collect two auxiliary human datasets that bracket the alignment axis. Both reuse the in-domain human data collection setup above; only the demonstrator's execution strategy and viewpoint change.

- **Unaligned (`bag-grocery`).** A counterfactual set in which the demonstrator performs the task in a manner whose retargeted trajectory yields *inexecutable* robot actions, e.g., grasps the parallel-jaw gripper cannot reproduce (Fig. 6, right).
- **Aligned (`cup-on-saucer`).** The demonstrator deliberately mimics the robot's motion, with camera height and viewpoint matched to the robot's static Aria mount (Fig. 9). This isolates execution alignment from sensing alignment by collapsing the latter.

EgoVerse Dataset. Our large-scale, in-the-wild human regime is the EgoVerse-A flagship split [42], which spans diverse scenes, objects, and demonstrators with *no* scene or object alignment to the robot data. We use the per-task flagship subsets for `cup-on-saucer`, `fold-clothes`, and `bag-grocery`, giving an overall ∼10:1 human-to-robot ratio (Table III). Each demonstration carries 3D hand pose for both hands (21 keypoints per hand in the camera frame) paired with a calibrated 6-DoF head pose from visual–inertial SLAM, which we use both to align the action channel and to stabilize the 3D-flow target into the camera frame at t (Sec. B-D).

C. Rollout Evaluation Protocol

Experiment Methods. Each method is evaluated on **ID** (20 rollouts: seen objects and scene, randomized positions and orientations) and **OOD** (20 rollouts: 10 unseen objects in the training scene, 10 seen objects in novel scenes with varied backgrounds and table heights), for **1800** rollouts in total across all methods and tasks. For every task we define task-specific sub-task metrics (grasps, placements, intermediate manipulations, and full completion) and report a normalized

sub-task score aggregated across rollouts alongside the binary success rate. Initial conditions are randomized and held common across methods within each task to ensure a fair comparison.

Cup-on-Saucer. The robot must reorient a cup from a randomized initial pose and place it upright on a saucer at a randomized position, demanding precise bimanual regrasping and fine-grained transport. Sub-task credit (1 point each, 3 total) is assigned for (i) rotating and picking up the cup, (ii) a successful handover between the two arms, and (iii) placing the cup upright on the saucer. SR is the fraction of rollouts in which the cup is correctly placed on the saucer.

Fold-Clothes. The robot must three-fold a T-shirt initialized in random configurations, a deformable task with shape variation and self-occlusion across stages. Sub-task credit (1 point each, 3 total) is assigned for (i) the bottom-sleeves fold, (ii) the top-sleeves fold, and (iii) the final fold in half. SR requires all three stages to complete cleanly.

Bag-Grocery. The robot must open a grocery bag and load three items into it from randomized positions—a long-horizon task in which the two arms first grasp the handles to open the bag, then insert items one by one. Sub-task credit is 1 point for opening the bag and 1 point per item placed inside, for a maximum of 4 points per rollout; an insertion counts only if the objects are placed in left-to-right order inside the bag. SR requires all three items loaded under the ordering constraint. Because its pick-and-place motions are naturally aligned between human and robot, this is the one task where action-level co-training helps, which we exploit in Sec. IV(**Q3**).

F. Simulation Experiments

Our main results are on real-world bimanual tasks (Sec. IV), where the human–robot embodiment gap is the object of study. To complement them with a *public, reproducible* testbed, we additionally instantiate EgoWAM on the LIBERO simulation benchmark [30]. The simulation study serves three purposes: (i) it validates that the single-backbone / swappable-world-head architecture (Sec. B-C) trains and evaluates end-to-end in a standard benchmark, (ii) it confirms that adding the auxiliary world-model channel does *not* degrade single-embodiment policy learning relative to a behavior-cloning policy, and (iii) it exercises the full world-target pipeline—Pixel, DINO, and 3D flow—in closed loop. LIBERO is a single-arm Franka setup with no human data, so it is ***not*** a test of cross-embodiment transfer; that claim rests on the real-world experiments.

A. Benchmark and Task Setup

LIBERO is built on robosuite [67] and MuJoCo, with a single 7-DoF Franka Panda and a parallel-jaw gripper under an operational-space (OSC) end-effector controller at $\sim$20 Hz. Actions are 7-dimensional per-step deltas $[\Delta x\ \Delta y\ \Delta z\ \Delta r_x\ \Delta r_y\ \Delta r_z\ \ g]$, where the pose terms are end-effector deltas in the controller frame and $g \in [-1\ 1]$ is the gripper command ($+1$ close, -1 open). Observations are a front-view and a wrist (eye-in-hand) camera, both rendered at native 128×128, together with end-effector pose and gripper proprioception. We evaluate on two task suites, each with 10 tasks and 50 demonstrations per task (500 demonstrations per suite): `libero-object`, which varies the manipulated object under a fixed scene, and `libero-spatial`, which varies object placement and spatial relations. Each task carries fixed initial states for deterministic benchmarking.

B. Integration into EgoWAM

We add LIBERO as a new embodiment to the heterogeneous tokenizer stack, reusing the shared trunk and flow-matching action head unchanged.

Action and Coordinate Conventions. Unlike the real-world bimanual setup, which predicts absolute end-effector pose chunks in a camera-aligned frame, the policy here predicts LIBERO's native 7-D per-step end-effector deltas directly, matching the convention used by every LIBERO baseline and making success numbers comparable. Absolute end-effector pose is retained only as a proprioceptive *observation*. Because LIBERO is a standalone single embodiment, we train in the robot base/world frame with identity extrinsics and omit the cross-embodiment camera-frame alignment used for human–robot co-training (Sec. B-A). The gripper is supervised on the recorded delta command and clamped to $[-1\ 1]$ at evaluation.

Simulator Isolation. LIBERO's simulator stack has package dependencies that conflict with our training environment. We therefore run the simulator in an isolated process that exchanges observations and actions with the policy over a lightweight interface. This keeps the two environments separate and lets us evaluate in closed loop directly from a training checkpoint, with no change to the model.

C. World-Model Variants and Training

All variants share the trunk and action head and differ only in the world-model target, mirroring the controlled design of Sec. B-D:

- **BC** — action-only policy with no world-model head.
- **Pixel** — future-frame prediction in the latent space of WAN VAE encoder [46].
- **DINO** — future DINO patch-feature prediction through the RAE wide-DDT head [38, 62].
- **3D Flow** — dense 3D point-flow prediction with targets from Track4World [32]. Since the `agentview` camera is static, ego-motion stabilization reduces to the identity, so the flow target reflects object motion directly.

The trunk is the same HPT backbone (256-d, 16 blocks, 8 heads, one-frame observation history) with an action horizon of 16 and a future horizon of 16; the world head is supervised under the shared flow-matching path with weight $\lambda = 1$. Each suite is trained multi-task with a per-episode one-hot task vector routed through the proprioception path. We train for 1000 epochs of 100 steps with AdamW (learning rate 3×10^{-4}, cosine annealing) in bf16 on a single NVIDIA L40S GPU, batch size 64 (BC/Pixel/DINO) or 32 (3D flow).

D. Closed-Loop Evaluation Protocol

Metric. The primary metric is closed-loop task success rate in the LIBERO MuJoCo simulator: a trial succeeds if the environment reports success at any point within the step budget (sparse success).

Faithfulness to LIBERO's native evaluation. Our protocol reproduces LIBERO's official evaluation loop, changing *only* the action interface (below). We run 20 parallel environments with 20 trials per task. Our initialization is LIBERO-identical: the policy starts each trial from the exact same fixed initial states that LIBERO's official evaluator uses. After each reset we take five zero-action settling steps before the policy acts, cap each rollout at 600 steps, and use a single fixed random seed throughout. In short, we keep LIBERO's native loop—20 environments $\times$ 20 trials, fixed initial states, 600-step horizon, sparse success—and change only the action interface.

Action interface (the one intentional deviation). LIBERO's reference policies emit one action per control step. Our flow-matching policy instead predicts a 16-step chunk of per-step end-effector deltas (no interpolation) and executes it with a receding horizon—applying the first 8 actions, then replanning—with actions clipped to $[-1\ 1]$. Everything else matches the native loop.

Fair comparison. Because training uses a fixed number of optimizer steps per epoch (independent of dataset size), the BC baseline and all world-model variants are trained to the same epoch budget (1000) and compared under identical training, evaluation, and seed.

Peak reporting. We checkpoint every 100 epochs and evaluate every checkpoint with the identical 20-trial protocol, reporting each method's *peak* success rate (and the full success-vs-epoch curve) rather than only the final cosine-annealed checkpoint; peaks are essentially equal to the final on `libero-object` but run up to $\sim$5% above it on `libero-spatial`, and often occur before the last epoch. A training-time monitor runs this same evaluation routine every 100 epochs, so the live training curve is methodologically identical to the reported numbers.

Reproducibility. Evaluation is reproducible run-to-run: we fix all random seeds and enable deterministic GPU computation. The simulator itself is deterministic given identical actions, so the policy is the only potential source of nondeterminism; two independent runs produce identical per-task results.

E. Results

Table IV reports peak closed-loop success on the two suites (20 trials per task, 200 rollouts per suite). In-domain success is near-saturated throughout, and *every* world-model variant matches or exceeds the BC baseline on both suites—confirming that the swappable world-model heads integrate cleanly and that the auxiliary world channel does not harm single-embodiment policy learning. On `libero-object` all variants sit at or above 96 5%, with Pixel saturating at 100%; on the harder `libero-spatial` the world-model variants lead BC by 2–7 points, the DINO target highest at 93 5%. We read these results as an architecture- and pipeline-level validation rather than a ranking of world representations—LIBERO is single-embodiment with no human data, so it does not exercise the cross-embodiment transfer the world target is designed for. The discriminative comparison among Pixel, DINO, and 3D flow is the role of the real-world cross-embodiment evaluation in Sec. IV.

TABLE IV: In-domain peak closed-loop success rate (%) on LIBERO (200 rollouts per suite, single seed). Variants share the trunk and action head, differing only in the world target; best per suite in **bold**.

Variant	`libero-object`	`libero-spatial`
BC	96.5	86.0
Pixel-WAM	**100.0**	90.5
DINO-WAM	97.0	**93.5**
3D Flow-WAM	99.0	88.0