EGO4WAM: WHAT MATTERS WHEN SCALING EGOCENTRIC HUMAN DATA FOR ROBOT LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric human data provides a scalable source of experience for robot learning, but varies substantially in human-robot alignment, behavioral coverage, and available supervision. Existing work shows favorable scaling with increasing human data, but it remains unclear which data properties drive downstream robot gains and how to use such data throughout the training pipeline. We present a systematic study of egocentric human data with different alignment and supervision under a unified world-action model framework. With the model backbone fixed, we disentangle the effects of human-robot alignment, data duration and task diversity, action supervision, and data usage strategies. We find that aligned human demonstrations substantially improve out-of-distribution generalization and reduce target-task robot data requirements; data duration and task diversity affect downstream capabilities differently; and video-only supervision remains effective without action labels, providing a strong foundation for subsequent video-action training. We validate these findings through closed-loop policy evaluation on both real robots and RoboDojo. Rather than treating data duration as the sole scaling axis, Ego4WAM shows how alignment, task diversity, available supervision, and usage strategy jointly shape the value of egocentric human data for robot learning.

1 INTRODUCTION

Egocentric human data offers a scalable source of experience for robot learning, covering a breadth of objects, scenes, and task variations that is difficult to match through robot teleoperation alone (Grauman et al., 2022; Damen et al., 2022; Hoque et al., 2026). Such experience can provide different forms of supervision for robot learning. Human hand actions provide action supervision for imitation learning (Kareer et al., 2025b; Li et al., 2025), while World-Action Models (WAMs) (Li et al., 2026a) additionally leverage future-state prediction as world-modeling supervision to learn task-relevant dynamics. Recent studies have shown strong human-to-robot transfer and favorable scaling with increasing egocentric data (Intelligence et al., 2025; Zheng et al., 2026; Team, 2026; DynaRobotics, 2026), yet it remains unclear which data properties drive downstream robot gains and how to use the data throughout the training pipeline. This raises a central question: **When scaling egocentric human data for robot learning, which data properties and usage strategies matter most?**

Egocentric data varies along several dimensions that directly affect its utility for robot learning. At one end of the spectrum, human demonstrations can be carefully aligned with robot data in viewpoint, motion speed, and behavior style, facilitating cross-embodiment learning but limiting the diversity of experience that can be collected. Relaxing these alignment constraints allows egocentric datasets to cover a much wider range of tasks, objects, and environments. However, obtaining reliable human action trajectories requires accurate camera-pose and hand-pose estimation, introducing additional processing cost and reducing label reliability (Li et al., 2026b; Zheng et al., 2026). Nevertheless, videos without action labels still contain informative interaction dynamics that WAMs can exploit through world modeling (Kareer et al., 2025a; Li et al., 2026a; DynaRobotics, 2026). These factors are often entangled when egocentric data is summarized only by duration, making it difficult to determine whether downstream gains arise from stronger human-robot alignment, broader behavioral coverage, different training objectives, or increased data scale itself.

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

To disentangle these factors, we conduct a systematic study of how to scale and use egocentric human data for robot learning under a unified WAM framework. We keep the model backbone fixed across our main experiments, while varying data alignment, coverage, supervision, and usage strategy. We first study human-robot alignment by examining whether aligned human demonstrations can improve out-of-distribution robot generalization and reduce the need for robot data. We then separate data duration from task diversity, comparing denser coverage of existing tasks with broader coverage across different behaviors. Finally, we relax action supervision and investigate whether large-scale egocentric videos without reliable action labels can still contribute through world modeling, and how such supervision interacts with subsequent video-action training. In this way, we disentangle data properties and usage strategies that aggregate scaling otherwise conflates.

Crucially, we evaluate the utility of egocentric data by measuring its effect on closed-loop policy performance on both real-world and RoboDojo (Chen et al., 2026) manipulation tasks, rather than relying on human-action prediction MSE alone. Across these studies, we find that egocentric data provides distinct benefits under different conditions. Human demonstrations covering object and scene conditions absent from robot training substantially improve OOD generalization and reduce the amount of target-task robot data required. Increasing data duration and task diversity leads to different performance trends across downstream capabilities. Video-only experience remains valuable beyond the subset with reliable action supervision, providing a strong foundation for subsequent video-action training. These findings show that the value of egocentric data depends not on scale alone, but on the distributions it covers, the supervision it reliably provides, and how it is used alongside robot data across the training pipeline. Our contributions can be summarized as follows:

- We study the utility of egocentric human data for robot learning under a controlled WAM framework. With the WAM backbone fixed, we disentangle data properties and usage strategies that large-scale scaling studies often couple.
- We characterize how human-robot alignment, data duration and task diversity, and available supervision affect downstream robot learning. This reveals distinct roles of these factors that are obscured when egocentric data is characterized only by its overall scale.
- We evaluate these effects directly through closed-loop policy performance on real robots and simulation rather than relying on human-action prediction metrics alone. Our results provide practical guidance on what egocentric data to collect, what supervision to obtain, and how to use human data across robot learning.

2 Related Works

Egocentric Human Data. Egocentric human data has attracted increasing attention in robot learning because it can be collected at substantially lower cost than robot teleoperation while covering a much broader range of objects, scenes, tasks, and interaction patterns. Large-scale datasets such as Ego4D (Grauman et al., 2022), Egocentric-10K (BuildAI, 2025), and RekaDaily-10K (RekaAI, 2026) primarily capture first-person activities from everyday life and work, providing large-scale visual and behavioral experience. More recent manipulation-oriented datasets, including EgoDex (Hoque et al., 2026), EgoLive (Li et al., 2026c), and EgoSuite (LightwheelAI, 2026), further augment egocentric observations with structured hand pose annotations (Pavlakos et al., 2024). Such action annotations make human experience more directly usable for robot policy learning, but obtaining reliable action trajectories introduces additional sensing, reconstruction, and quality-control requirements, limiting the amount of data that can be annotated with high confidence. A further step is to collect human demonstrations that are better aligned with downstream robot tasks. For example, EgoVerse (Punamiya et al., 2026) considers task-aligned and environment-aligned human data for human-to-robot transfer. Although stronger alignment facilitates transfer to robot learning, it also imposes stricter collection requirements. Consequently, modern egocentric data naturally spans a spectrum from large-scale video-only experience to action-annotated data and robot-aligned demonstrations, with different levels of alignment and supervision. As illustrated in Figure 1, we organize this spectrum as an *egocentric data pyramid*, retaining each sample at the most reliable supervision level it can support. This perspective provides a basis for studying which data properties deserve more attention, what supervision is worth obtaining, and how to use egocentric data as its scale continues to grow.

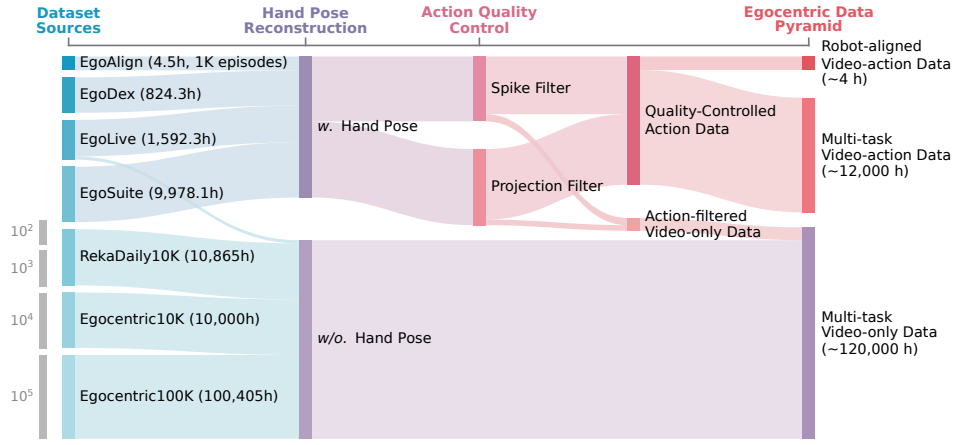


Figure 1: Overview of the egocentric data pyramid.

Utilizing Egocentric Data for Robot Learning. Recent work increasingly studies how egocentric human experience can support robot learning across different training settings. For task-aligned transfer, EgoVerse (Punamiya et al., 2026) shows that broad human experience can be adapted effectively with a small amount of aligned human–robot data, while EgoWAM (Li et al., 2026a) investigates how different world-prediction objectives transfer human experience under controlled human–robot alignment. At larger scale, Being-H0 (Luo et al., 2025) and EgoScale (Zheng et al., 2026) use retargeted human actions to pretrain transferable manipulation priors, while Being-H0.7 (Team, 2026) and OpenWAM (Wang et al., 2026c) further introduce future or video prediction so that egocentric observations can contribute through world-model supervision. Complementary efforts reduce the embodiment gap by explicitly transforming human experience toward the robot domain through hand and wrist retargeting, action-space or viewpoint alignment, and visual embodiment editing (Wang et al., 2026b). Such conversion can make human data easier to incorporate into robot policies, but also introduces additional dependence on camera tracking, hand reconstruction, and visual synthesis, whose errors and embodiment ambiguities can limit reliability. More recently, EgoScale (Zheng et al., 2026), Motus2 (Bi et al., 2026), and Dyna-2 (DynaRobotics, 2026) have demonstrated favorable scaling trends with increasingly large human-data mixtures. These results establish the scalability of human experience, but aggregate scaling curves and open-loop human prediction metrics do not fully reveal which properties of egocentric data drive downstream robot gains, or whether their relative benefits persist under closed-loop deployment. Rather than focusing on converting human data into robot-compatible supervision, we systematically study data properties and usage strategies under a controlled WAM framework. We evaluate these factors directly through closed-loop robot performance, aiming to provide actionable guidance on what egocentric data to collect, what supervision to retain, and how to use such data for robot learning.

3 Method

Our goal is to provide a controlled framework for studying how egocentric data contributes to robot learning under different supervision and alignment settings. We therefore keep the model backbone and unified action space fixed across our main experiments, while varying the data composition, supervision, and usage strategies.

3.1 Egocentric Data Pyramid

As discussed in Section 2 and illustrated in Figure 1, egocentric data varies substantially in human-robot alignment and available supervision. We organize these differences into an egocentric human data pyramid, ranging from a small amount of robot-aligned human demonstrations to large-scale action-annotated data and a larger collection of video-only data.

EgoAlign is our internally collected dataset, processed separately to form the robot-aligned subset with closely matched task and behavioral distributions. We obtain human actions through a hand-

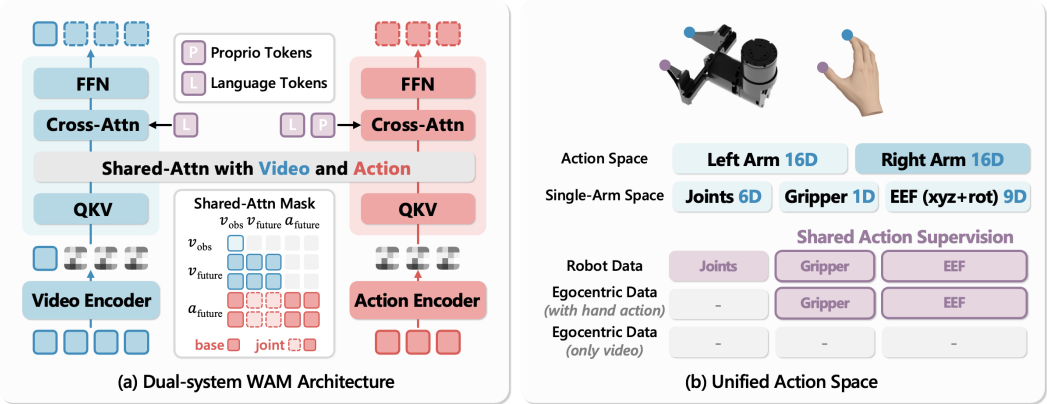


Figure 2: Illustration of the world-action model framework and unified human-robot action space.

pose and camera-pose annotation pipeline, with implementation details in the appendix. For public datasets with available action annotations, we further apply temporal and geometric filters to control action quality, with the detailed filtering criteria described in the appendix. Samples that pass these filters form our quality-controlled multi-task video-action data. We retain samples without action annotations, along with those rejected by the action-quality pipeline, as video-only data whenever the visual interaction remains valid. Although they do not provide reliable action supervision, they still capture object motion, interaction outcomes, and scene transitions that can support world modeling. Together, these data constitute the egocentric data collection from which we construct the training subsets used in our subsequent experiments.

3.2 World-Action Model Framework

Building on Yuan et al. (2026), we design the backbone used throughout our study, as illustrated in Figure 2 (a). The model consists of a video branch for world modeling and an action branch for action prediction. Language instructions are used as conditioning signals, while proprioceptive states are provided only to the action branch and are not injected into the world-modeling branch. The two branches interact through shared attention. We consider two shared-attention mechanisms in our study. In the *base* formulation, future action tokens attend only to the current visual context. In the *joint* formulation, they can additionally attend to predicted future visual representations, while future visual tokens remain independent of action tokens in both settings. This formulation naturally supports different supervision regimes. Video-action data can jointly supervise world and action prediction, while video-only data can independently supervise world modeling.

3.3 Unified Action Space

Human and robot actions differ substantially in their embodiment-specific control dimensions. Rather than fully retargeting human motion to robot joint commands, we define a unified action space that explicitly separates robot-specific control from components shared across embodiments. As shown in Figure 2 (b), all tasks in our study use bimanual control. Each arm is represented by a 16D action vector consisting of a 6D joint signal, a 1D gripper state, and a 9D end-effector (EEF) representation, where the EEF component contains 3D translation and a 6D rotation representation. We concatenate the two arms' actions into a 32D bimanual action space.

For egocentric human data with action annotations, we supervise only the components that admit a meaningful cross-embodiment correspondence. Following prior human-robot co-training practice, we use only the EEF pose and gripper state as the default shared representation for human motion. To handle the missing dimensions in the flow-matching action space, we use the analytic forward-noise strategy used in Wang et al. (2026c), and provide the detailed formulation in the appendix.

3.4 Training Protocol

We organize training into three stages according to the supervision available in our egocentric data pyramid. We use *pre-training*, *mid-training*, and *post-training* to denote world-model pre-training on egocentric video data, joint world-action mid-training on action-annotated data, and downstream robot adaptation, respectively. During pre-training, we optimize only the world-modeling objective on egocentric video. Mid-training uses samples with reliable actions and jointly optimizes world and action prediction. Post-training adapts the model to downstream robot manipulation using embodiment-specific robot demonstrations, optionally together with robot-aligned human data. We keep the model architecture fixed across these stages unless otherwise specified. The appendix provides detailed data mixtures, optimization schedules, and implementation settings.

4 Experiments

Our experiments follow the egocentric data pyramid, progressively relaxing constraints on human-data scaling. We begin with robot-aligned demonstrations to study OOD generalization and robustness. We then move to large-scale multi-task video-action data to examine transferable manipulation priors and scaling along data duration and task diversity. Finally, we relax action supervision to investigate whether video-only experience can further benefit robot learning.

We evaluate the effect of egocentric data on downstream robot performance through closed-loop evaluation, rather than relying on human-action prediction MSE as an open-loop proxy. For real-world evaluation, we use a bimanual Piper robot with parallel grippers and collect robot demonstrations through teleoperation. The appendix provides detailed task definitions, data collection protocols, and ID and OOD evaluation settings. For broader evaluation, we follow the official RoboDojo (Chen et al., 2026) protocol and report both progress score and success rate. Since our model does not explicitly target long-term memory or VLM-based open instruction following, we focus primarily on *Generalization*, *Precision*, and *Long-Horizon*, while reporting results across all benchmark categories for completeness.

4.1 Robot-Aligned Human Demonstrations

We first study robot-aligned human demonstrations to examine two questions: whether they improve out-of-distribution (OOD) generalization and robustness to unseen object and scene variations, and whether they reduce the amount of target-task robot supervision required. In particular, we conduct real-robot experiments on *Place into the Basket* and *Fold Cloth*, representing rigid-object pick-and-place and deformable-object manipulation, respectively. For each task, we collect 300 in-domain (ID) robot demonstrations and a matched human set containing 300 ID, 100 object-OOD, and 100 scene-OOD demonstrations under the same task specification. The OOD conditions are absent from robot training and are covered only by the corresponding human demonstrations.

Generalizing to Environments Unseen in Robot Data. Figures 3 and 4 show that human demonstrations from object or scene conditions absent from robot training can substantially improve robot performance under the corresponding conditions. On *Place into the Basket*, the robot-only policy achieves 60% ID success but drops to 10% on object-OOD and 0% on scene-OOD. Introducing object-OOD human demonstrations raises object-OOD success to 60%, and introducing scene-OOD human demonstrations raises scene-OOD success to 20% with a substantially larger gain in task progress. These results show that the model can acquire task-relevant knowledge about new objects and environments directly from human demonstrations. This provides a practical way to extend a robot policy to new objects and scenes without requiring corresponding robot demonstrations.

Reducing Dependence on Robot Data. We next examine whether aligned human demonstrations can reduce the amount of target-task robot data required for policy learning. We vary the number of robot demonstrations on *Fold Cloth* while keeping the same 500 human demonstrations fixed. Reducing robot data from 300 to 100 demonstrations causes a substantial performance drop, whereas adding human data recovers 80% ID, 60% object-OOD, and 80% scene-OOD success. However, when robot supervision is further reduced to only 20 demonstrations, the same human data is no longer sufficient to recover effective policy performance. These results show that aligned human

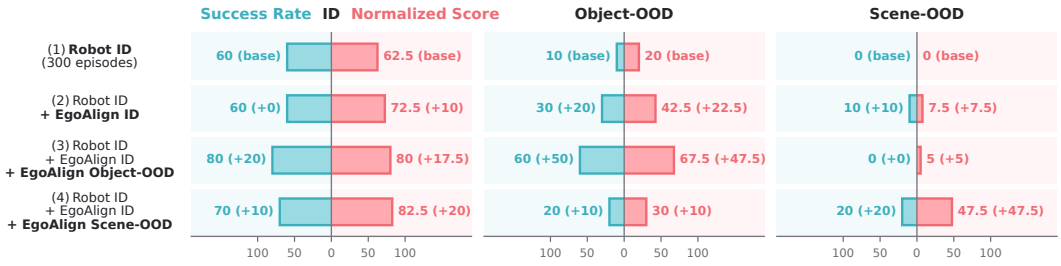


Figure 3: Real-robot evaluation on *Place into the Basket* with human data under OOD settings.

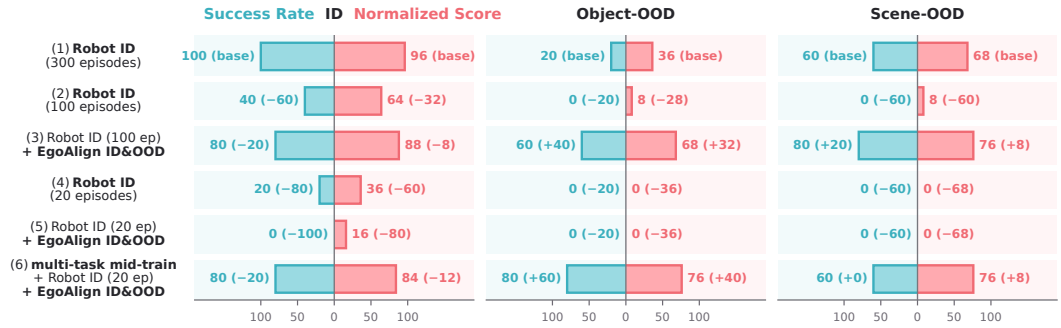


Figure 4: Real-robot evaluation on *Fold Cloth* with human data under different robot-data budgets.

demonstrations can substantially reduce the dependence on target-task robot data, but cannot fully replace embodiment-specific robot supervision.

4.2 Scaling Beyond Task-Specific Egocentric Data

Although robot-aligned human demonstrations can effectively help models adapt to specific distributions, their dependence on target-task collection limits scalability. We therefore move to large-scale multi-task egocentric data collected independently of downstream tasks and investigate whether this experience can establish transferable manipulation priors before downstream adaptation. We further study how scaling along two key dimensions, data duration and task diversity, affects downstream robot learning. We call this stage *multi-task mid-training*, where we optimize both video and action objectives. In addition to egocentric human data, we include approximately 90 hours of real-robot data to provide embodiment-specific grounding in the robot action space. The robot data contains only two generic task families, pick-and-place and towel folding, and the corresponding tasks, objects, and scenes do not overlap with downstream evaluation.

Further Reducing Dependence on Robot Data. As shown in Figure 4, with only 20 target-task robot demonstrations, adding the full aligned human set alone fails to recover useful performance. In contrast, after multi-task mid-training, the model achieves 80% success on ID, 80% on object-OOD, and 60% on scene-OOD using the same 20 robot demonstrations and aligned human data. Importantly, the amount of target-task robot data and the aligned human data do not change between these settings. The improvement therefore reflects the transferable priors acquired during multi-task mid-training. This suggests that broad multi-task experience can make limited downstream data more effective, further reducing dependence on target-task robot data.

Scaling Data Duration and Task Diversity. We next study how to scale multi-task human data by decomposing data scale into two dimensions: the data duration within a fixed task set and the diversity of tasks covered. We construct two controlled scaling sequences from the semantic task categories of our egocentric data. For *duration scaling*, we fix the same 500 most frequent tasks and increase the total duration from 1K to 2K and 4K hours. For *task scaling*, we progressively expand from 500 to 1K, 2K, and 6K tasks toward the naturally occurring long tail, while keeping the average duration per task approximately constant, resulting in 1K, 2K, 4K, and 12K hours of data.

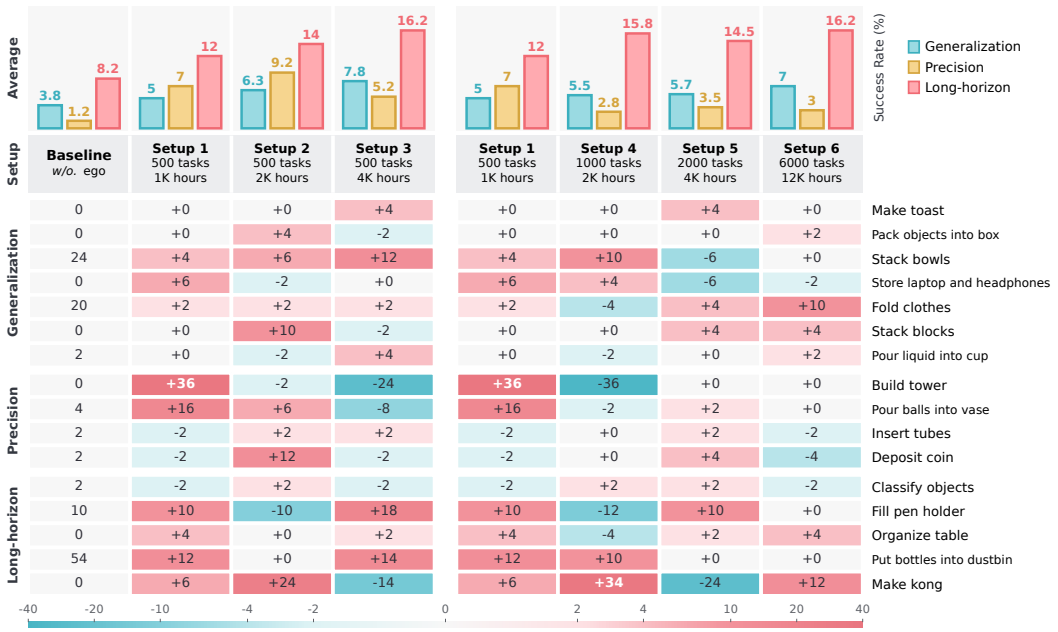


Figure 5: **RoboDojo evaluation under data-duration and task-diversity scaling.** We visualize a subset of tasks with non-zero success rate, while all aggregate results are computed over the full evaluation task set. We provide complete per-task results in the appendix.

As shown in Fig. 5, the two scaling directions lead to different capability profiles. With a fixed task set, increasing duration from 1K to 4K hours improves Generalization from 5.0 to 7.8 and Long-Horizon from 12.0 to 16.2. Increasing task diversity from 500 to 6K tasks similarly improves Generalization from 5.0 to 7.0 and Long-Horizon from 12.0 to 16.2, but Precision decreases as we introduce increasingly long-tail tasks. At matched data volumes of 2K and 4K hours, denser coverage of the top 500 tasks also remains stronger on several aggregate metrics than broader task sets. The task-level results further reveal substantial positive and negative transfer.

These results show that data duration and task diversity are distinct, capability-dependent scaling axes. More data within existing tasks and broader task coverage benefit different capabilities and can also introduce trade-offs. Consequently, neither total hours nor task count alone adequately characterizes useful egocentric-data scale. The results suggest ensuring sufficient coverage of common behaviors before expanding aggressively into increasingly sparse long-tail tasks.

4.3 Relaxing Action Supervision with Large-Scale Human Video

As egocentric data scales, obtaining reliable action supervision becomes increasingly costly. Human action labels rely on accurate hand reconstruction, camera tracking, and temporal quality control, while reconstruction failures and noisy trajectories further reduce usable video-action data. In contrast, the corresponding videos can still preserve informative object motion, interaction outcomes, and physical state transitions. Since WAMs can learn such dynamics directly through video prediction, we investigate whether video-only supervision can improve downstream robot learning without requiring reliable action labels, and how its benefits interact with subsequent video-action mid-training.

We call this stage *video pre-training*, where we optimize only the video-prediction objective. The pre-training set contains the same 12K hours of action-valid egocentric videos used for mid-training, plus approximately 3K additional hours sampled from the video-only data, for a total of 15K hours. Mid-training subsequently uses the 12K-hour action-valid subset together with the multi-task robot data described in Section 4.2, jointly optimizing video and action objectives. The two stages are not designed as a compute-matched comparison, since video-only data is naturally available at a larger

Table 1: **Ablation study of egocentric data utilization strategies on RoboDojo.** Base and joint denote different shared-attention mechanisms used during post-training.

Training Setup			Genera.		Precision		Long		Memory		Open		Average	
pre-train	mid-train	post-train	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR
-	-	base	5.87	3.83	5.56	1.25	16.97	8.25	2.75	1.67	0.78	0.75	6.39	3.15
-	✓	base	10.73	7.00	8.45	3.00	25.18	16.25	6.38	5.33	0.68	0.50	10.28	6.42
✓	-	base	12.04	7.67	23.53	16.25	27.85	17.50	6.67	5.33	0.55	0.50	14.13	9.45
✓	✓	base	12.63	8.33	24.77	18.00	29.46	20.50	7.42	6.33	0.90	0.75	15.04	10.78
✓	-	joint	17.19	12.17	31.16	22.25	44.16	29.75	8.37	7.33	0.55	0.50	20.29	14.35

Table 2: **Performance comparison on RoboDojo.** The scale of robot and egocentric training data is reported in hours. Results of existing methods are taken from the official leaderboard. The style of **bold** and underline denotes the best and second-best results, respectively.

Model	Data Scale		Genera.		Precision		Long		Memory		Open		Average	
	Robot	Ego	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR	Score	SR
Pi-0.5	-	-	13.38	8.17	12.40	5.50	23.54	14.67	5.89	4.67	1.98	1.67	11.44	6.93
Spatial Forcing	-	-	14.12	9.34	17.32	10.58	23.26	14.58	5.43	4.11	1.78	1.58	12.38	8.04
Hy-0.5-VLA	-	-	11.78	8.39	13.81	8.00	25.74	14.92	13.37	12.11	0.65	0.58	13.07	8.80
Meituan-0	-	-	13.75	8.17	16.77	7.75	29.61	18.58	10.06	8.89	4.54	4.25	14.95	9.53
Xiaomi-1	7.2K	100K	23.54	17.00	26.69	18.83	38.39	23.67	7.81	6.56	3.94	3.58	20.07	13.93
GalaxeaVLA-0.5	-	-	18.46	12.83	28.25	20.42	44.12	**32.25**	8.61	7.33	1.73	1.58	20.23	14.88
GPT-6-Astra	-	-	**33.36**	**30.50**	12.65	4.00	21.45	8.25	**43.04**	**38.67**	**34.36**	**31.00**	**28.97**	**22.48**
FastWAM*	*w/o.* pretrain		5.87	3.83	5.56	1.25	16.97	8.25	2.75	1.67	0.78	0.75	6.39	3.15
OpenWAM-α	4.9K	1.4K	20.71	14.83	18.45	9.25	34.93	25.33	10.41	9.11	1.41	1.08	17.18	11.92
Ego4WAM	0.1K	15K	17.19	12.17	**31.16**	**22.25**	**44.16**	29.75	8.37	7.33	0.55	0.50	20.29	14.35

scale than reliably action-annotated data. This scale difference is reflected in the egocentric data pyramid and motivates studying how the two supervision regimes can be used together.

Learning Dynamics Priors without Action Supervision. As shown in Table 1, the model without human pre-training or mid-training achieves an average RoboDojo score of 6.39 with a 3.15% success rate. Video-only pre-training improves this to 14.13 / 9.45% despite using no human action supervision, showing that egocentric experience remains valuable even when reliable action labels are unavailable. WAMs can therefore extend the usable data beyond the subset that supports imitation learning.

Applying video-action mid-training after pre-training further improves performance from 14.13 / 9.45% to 15.04 / 10.78%. This relatively small gain suggests that the two stages capture partially overlapping information about manipulation dynamics. More importantly, the same mid-training without prior video pre-training reaches only 10.28 / 6.42%. The substantial gap between these two settings highlights the importance of first establishing a strong world-modeling branch for WAMs, so that subsequent action learning can build on informative visual dynamics. Egocentric video is particularly suitable for this role, as it can be scaled without action labels while retaining dynamics.

Leveraging Dynamics Priors for Joint World-Action Learning. The results above show that large-scale video pre-training provides a strong dynamics prior for the world-modeling branch. We next ask whether this improved world representation can be more effectively exploited during downstream action learning. We compare the *base* and *joint* shared-attention mechanisms introduced in Section 3.2, where the joint formulation additionally allows action prediction to access predicted future visual representations. Starting from the same video-pretrained model, switching from base to joint substantially improves the average RoboDojo score from 14.13 to 20.29 and the success rate from 9.45% to 14.35%. This suggests that stronger dynamics priors make predicted future representations more informative for action generation, allowing joint world-action learning to better exploit the knowledge acquired from egocentric video.

Interestingly, FastWAM (Yuan et al., 2026) reports only small differences between its base and joint variants without embodied pre-training. Our results suggest a complementary explanation that the



Figure 6: Real-robot evaluation across diverse manipulation tasks.

benefit of joint future conditioning depends on the quality of the learned world representation. With large-scale egocentric video pre-training, the same basic WAM backbone achieves an average RoboDojo score of 20.29 and a 14.35% success rate using only approximately 0.1K hours of multi-task robot data and 15K hours of egocentric experience. This result further demonstrates the potential of egocentric data to strengthen world dynamics in WAMs and highlights the importance of studying not only how much human data is available, but also how it is used throughout the training pipeline.

4.4 More Results on the Real Robot Tasks

Finally, we evaluate on a broader set of real-world manipulation tasks beyond the settings used above. As shown in Figure 6, we consider tasks spanning rigid-object manipulation, deformable-object interaction, and longer-horizon composition: *Fold Towel*, *Tighten a Bottle Cap*, *Stack Plates and Bowls*, and *Heat Food with a Microwave*. We provide videos on the project page.

5 Conclusion

We presented a systematic study on scaling egocentric human data for robot learning with WAM. Rather than treating the duration of human data as a single measure of scale, we disentangled the roles of human-robot alignment, data duration and task diversity, available supervision, and usage strategies. The study leads to several key **takeaways** for scaling egocentric human data:

- **Human data reduces the need for robot data, but cannot fully replace it.** Broad multi-task human experience provides transferable manipulation priors, aligned human demonstrations cover distribution variations, and a small amount of robot data remains important for embodiment adaptation. Their combination can substantially reduce the amount of target-task robot data.
- **Scaling egocentric data does not always lead to better performance.** Data duration and task diversity represent different ways to scale, and total hours alone cannot characterize useful data scale. Denser coverage of recurring tasks provides limited additional gains, while introducing increasingly sparse long-tail tasks can even hurt precision-sensitive performance.
- **Egocentric data provides a strong dynamics prior for world-action learning.** Video-only pre-training can learn strong dynamics priors without action labels, extending usable data beyond the action subset. Such priors further improve subsequent world-action learning, showing that the value of human data depends not only on its supervision, but also on when and how it is used.

Ego4WAM provides a practical perspective on how to collect and utilize human data for robot learning. We hope this study serves as an empirical reference for designing future egocentric-data pipelines, and we will release our code, training recipes, and model checkpoints at different training stages to support further research.

STATEMENT

AI Use Statement. Generative AI tools were used to assist with language editing, manuscript organization, and improving the clarity of presentation. The authors developed, reviewed, and verified all research ideas, experimental design, data processing, analyses, and conclusions. The authors take full responsibility for the content of this paper.

Ethics Statement. Our study involves egocentric recordings of human manipulation collected for research purposes. All participants in our internally collected data provided informed consent for data collection and research use. For publicly available datasets, we follow their respective licenses, terms of use, and data-use requirements. We do not intentionally collect or release personally identifiable information, and we use the collected data solely for research on robot learning.

Reproducibility Statement. We provide detailed descriptions of the model framework, action space, training stages, and evaluation protocols in the main paper and appendix. To facilitate reproducibility and further research, we will release our code, training recipes, and model checkpoints from different pre-training and mid-training stages. We will also release the RoboDojo post-trained checkpoints and submit them to the official RoboDojo evaluation pipeline, allowing independent verification of our reported benchmark results and direct comparison under the standardized evaluation protocol.

REFERENCES

Hongzhe Bi, Zihao Zhou, Yihang Tang, Jingrui Pang, Shuhe Huang, Haitian Liu, Runqing Wang, Shuai Huang, Yichen Wang, Yiming Cheng, et al. Motus2: A self-evolving general world model for dexterous manipulation. *arXiv preprint arXiv:2608.30237*, 2026. 3

BuildAI. Egocentric-10k, 2025. URL https://huggingface.co/datasets/builddotai/Egocentric-10K. 2

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, et al. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026. 2, 5, 17

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 2022. 1

DynaRobotics. Dyna-2: A 1-million-hour scaling law for world-action models, 2026. URL https://www.dyna.co/dyna-2. 1, 3

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026. 1, 2

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. 1

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)*, 2025a. 1

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*, 2025b. 1

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. Egowam: World action models beyond pixels with in-the-wild egocentric human data. In *Robot World Models*, 2026a. 1, 3

Hao Li, Ganlong Zhao, Yufei Liu, Haotian Hou, Guoquan Ye, Tongyan Fang, Chunxiao Liu, Siyuan Huang, Jianbo Liu, Xiaogang Wang, et al. Ace-ego-0: Unifying egocentric human and robotic data for vla pretraining. *arXiv preprint arXiv:2606.17200*, 2026b. 1, 14

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025. 1

Yihang Li, Xuelong Wei, Jingzhou Luo, Yingjing Xiao, Yibo Bai, Guangyuan Zhou, Teng Zou, Chenguang Gui, Jiajun Wen, He Zhang, et al. Egolive: A large-scale egocentric dataset from real-world human tasks. *arXiv preprint arXiv:2604.23570*, 2026c. 2

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, et al. Open-aoe: An open egocentric manipulation dataset and toolchain for embodied learning. *arXiv preprint arXiv:2607.14183*, 2026d. 13

LightwheelAI. Egosuite, 2026. URL https://huggingface.co/collections/LightwheelAI/egosuite-open100k. 2

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. 13

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. 3

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 13

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. 2, 3

RekaAI. Rekadaily-10k, 2026. URL https://huggingface.co/datasets/RekaAI/RekaDaily-10k-raw. 2

BeingBeyond Team. Being-h0.7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026. 1, 3

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω. *arXiv preprint arXiv:2605.15195*, 2026a. 13

Ye Wang, Pei Lin, Xiong-Hui Chen, Haoqi Yuan, Zhixuan Liang, Yiyang Huang, Anzhe Chen, Zixing Lei, Jie Zhang, Tao Zhang, et al. Ego2robot: Scalable robot data synthesis from egocentric human data. *arXiv preprint arXiv:2608.02580*, 2026b. 3

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026c. 3, 4, 14

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. 4, 8, 14

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. 1, 3

A EGOCENTRIC DATA

A.1 EGOCENTRIC DATA PYRAMID

Figure 1 summarizes the supervision hierarchy and available scale of the egocentric data considered in our study. The pyramid characterizes a general property of egocentric data rather than the exact data mixture used in each experiment. As stronger supervision is required, fewer samples can reliably satisfy the corresponding annotation and alignment requirements. Robot-aligned demonstrations therefore occupy the smallest tier, followed by multi-task video-action data with reliable human actions, while video-only data can be retained at substantially larger scale.

We construct the pyramid based on the most reliable supervision each sample supports. We process EgoAlign separately as the robot-aligned subset. For public datasets with available hand or action annotations, we apply temporal and geometric quality-control filters to remove unreliable trajectories. Samples that pass the action-quality checks form the multi-task video-action data. Samples without reliable action annotations, including those rejected by the action-quality pipeline, can still be retained as video-only data when the underlying visual interaction remains valid. This way, unreliable action labels do not require discarding otherwise useful interaction videos.

Importantly, the approximately 120K hours of video-only data shown in Figure 1 represent the available data scale and are not all used in our experiments. Due to practical training constraints, we construct task-specific training subsets from this collection. In our main video pre-training experiments, we use the 12K-hour action-valid subset with video supervision only, plus about 3K additional hours sampled from the video-only data, for a total of 15K hours of egocentric video. We then use the 12K-hour action-valid subset for multi-task mid-training with both video and action supervision. Thus, the pyramid primarily shows how the amount of usable egocentric experience increases as supervision requirements are relaxed, while the corresponding experiments specify the exact data mixtures used at each training stage.

A.2 EGOALIGN

EgoAlign is our internally collected egocentric dataset for studying robot-aligned human demonstrations. We record human demonstrations using a monocular head-mounted camera under task and environment settings matching the real-robot experiments. All participants provided informed consent for data collection and research use. We recover human hand actions from the recorded videos, referring to the general processing pipeline of OpenAoE (Li et al., 2026d).

Specifically, we first estimate the camera trajectory using VGGT-Omega (Wang et al., 2026a) and take the coordinate system of the first camera frame as the reference frame for each sequence. We then use GroundingDINO (Liu et al., 2024) to localize hand regions and apply HaMeR (Pavlakos et al., 2024) to estimate the 3D hand poses independently for each frame. Combining the estimated hand pose with the recovered camera trajectory transforms the frame-wise hand predictions into a temporally shared coordinate system referenced to the first camera frame. The resulting trajectories provide the hand motion used to construct the shared end-effector and gripper supervision for human demonstrations. Finally, we inspect and filter the reconstructed trajectories to remove samples with unreliable hand localization, implausible hand motion, unstable camera tracking, or obvious temporal and geometric inconsistencies. We retain only sequences that pass this quality-control process as action-annotated EgoAlign demonstrations.

B METHOD

B.1 UNIFIED ACTION SPACE

Converting Human Hand Pose to Gripper EEF. To jointly train on human and robot actions, we abstract each human hand as a parallel-gripper end effector. Each hand is represented by a 9D end-effector pose together with a 1D gripper-openness signal. The 3D end-effector position is defined as the midpoint between the thumb and index fingertips,

$$\mathbf{p}_{\text{eef}} = \frac{1}{2}\left(\mathbf{p}_{\text{thumb}} + \mathbf{p}_{\text{index}}\right), \tag{1}$$

while gripper openness is defined by their Euclidean distance,

$$g = \|\mathbf{p}_{\text{thumb}} - \mathbf{p}_{\text{index}}\|_2 . \tag{2}$$

For hand orientation, we follow the hand-centric construction of ACE-Ego-0 (Li et al., 2026b). Let $\mathbf{p}_{\text{wrist}}$ denote the wrist position and let $\mathbf{p}_{\text{palm}}$ be the centroid of the index, middle, and ring fingertips. We construct an orthonormal hand frame as

$$\mathbf{x} = \frac{\mathbf{p}_{\text{palm}} - \mathbf{p}_{\text{wrist}}}{\|\mathbf{p}_{\text{palm}} - \mathbf{p}_{\text{wrist}}\|_2}, \qquad \mathbf{z} = \hat{\mathbf{n}}\left(\mathbf{p}_{\text{wrist}}, \mathbf{p}_{\text{thumb}}, \mathbf{p}_{\text{middle}}\right), \qquad \mathbf{y} = \mathbf{z} \times \mathbf{x}, \tag{3}$$

where $\hat{\mathbf{n}}(\cdot)$ denotes the unit normal of the corresponding hand plane, with its direction chosen consistently with the palm orientation. The resulting rotation matrix $\mathbf{R}_{\text{hand}} = [\mathbf{x}, \mathbf{y}, \mathbf{z}]$ is converted to the continuous 6D rotation representation by concatenating its first two columns. Together with the 3D position, this yields the 9D EEF representation used for human action supervision.

Robot actions additionally contain embodiment-specific joint signals. As described in the main paper, each robot arm is represented by 6D joint states, a 1D gripper state, and the same 9D EEF representation. Human demonstrations supervise only the shared EEF and gripper components, leaving robot-specific joint dimensions unavailable.

Analytic Noise Path. Following the unified action interface of OpenWAM (Wang et al., 2026c), we embed embodiment-specific actions into a fixed-width action space and associate each sample with a validity mask indicating which action dimensions are supervised. Robot demonstrations provide both joint and shared EEF supervision, whereas human demonstrations activate only the shared EEF and gripper dimensions. We denote the active and inactive dimension sets by $\mathcal{A}$ and $\mathcal{I}$, respectively.

We train the action branch with flow matching. Given a clean action $\mathbf{x}_0$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, and noise level σ, the forward trajectory is

$$\mathbf{x}_\sigma = (1-\sigma)\mathbf{x}_0 + \sigma\boldsymbol{\epsilon}, \qquad \mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{x}_0. \tag{4}$$

Only valid action dimensions contribute to the training objective,

$$\mathcal{L}_{\text{action}} = \mathbb{E}\left[w(\sigma) \frac{\sum_{t,d} M_{t,d} \left\| \mathbf{v}_\theta(\mathbf{x}_\sigma, \sigma)_{t,d} - \mathbf{v}^\star_{t,d} \right\|_2^2}{\max\left(\sum_{t,d} M_{t,d}, 1\right)} \right], \tag{5}$$

where $M_{t,d}$ denotes the validity mask over temporal and action dimensions.

Simply masking the loss leaves predictions on inactive dimensions unconstrained, which can accumulate arbitrary values during iterative flow sampling. We therefore keep inactive dimensions on their analytic forward-noise trajectory. Since $\mathbf{x}_0[\mathcal{I}] = 0$, their training marginal is $\mathbf{x}_\sigma[\mathcal{I}] = \sigma\boldsymbol{\epsilon}_{\mathcal{I}}$. At the first sampling step, we recover the corresponding latent noise and keep it fixed throughout sampling,

$$\hat{\boldsymbol{\epsilon}}_{\mathcal{I}} = \frac{\mathbf{x}_{\sigma_0}[\mathcal{I}]}{\sigma_0}, \qquad \mathbf{x}_{\sigma_i}[\mathcal{I}] = \sigma_i \hat{\boldsymbol{\epsilon}}_{\mathcal{I}}. \tag{6}$$

Thus, inactive dimensions remain distributed consistently with the training forward process without contributing supervision or being numerically integrated by the sampler.

B.2 Optimization Objectives

Following FastWAM (Yuan et al., 2026), we train both the world-modeling and action-prediction branches with conditional flow matching. Let $\mathbf{y}$ denote either future video latents or an action trajectory. Given Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and a noise level $\sigma \in (0, 1)$, we construct the noisy sample as

$$\mathbf{y}_\sigma = (1-\sigma)\mathbf{y} + \sigma\boldsymbol{\epsilon}, \tag{7}$$

with the corresponding velocity target

$$\mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{y}. \tag{8}$$

The generic flow-matching objective is

$$\mathcal{L}_{\text{FM}}(\mathbf{y}) = \mathbb{E}_{\mathbf{y},\boldsymbol{\epsilon},\sigma}\left[w(\sigma) \left\| \mathbf{v}_\theta(\mathbf{y}_\sigma, \sigma, \mathbf{c}) - \mathbf{v}^\star \right\|_2^2 \right], \tag{9}$$

Table 3: Training configurations used throughout our experiments. All experiments within the same configuration use identical optimization settings unless otherwise specified.

Parameter	Video Pre-train	Full Mid-train	Scaling Mid-train	Real-Robot Post-train	RoboDojo Post-train
Training Steps	150K	80K	40K	10K	40K
Nodes	8	8	4	1	2
Initial LR	2×10^{-5}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Final LR	1×10^{-5}	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Warmup Steps	2K	2K	2K	1K	2K
Shared Attention	Base	Base	Base	Base	Base / Joint

where $\mathbf{c}$ denotes the corresponding conditioning signals and $w(\sigma)$ is the timestep-dependent loss weight.

World-Modeling Objective. For world modeling, $\mathbf{y} = \mathbf{z}_{1:T}$ denotes the latent representation of future video frames encoded by the video VAE. The world-modeling loss is

$$\mathcal{L}_{\text{world}} = \mathcal{L}_{\text{FM}}(\mathbf{z}_{1:T}). \tag{10}$$

The world branch is conditioned on the current visual observation and language instruction. This objective applies to egocentric videos regardless of whether reliable action annotations are available.

Action-Prediction Objective. For action prediction, $\mathbf{y} = \mathbf{a}_{1:H}$ denotes an action chunk of horizon H. We optimize

$$\mathcal{L}_{\text{action}} = \mathcal{L}_{\text{FM}}(\mathbf{a}_{1:H}), \tag{11}$$

where the action branch is additionally conditioned on the available proprioceptive state. For human demonstrations, we evaluate the loss only over the shared EEF and gripper dimensions defined in Section B.1; we handle unavailable embodiment-specific dimensions using the validity mask and analytic noise path described above.

Training with Different Supervision. The overall objective for an action-annotated sample is

$$\mathcal{L} = \lambda_{\text{world}}\mathcal{L}_{\text{world}} + \lambda_{\text{action}}\mathcal{L}_{\text{action}}, \tag{12}$$

where λ_{world} and λ_{action} control the relative contribution of the two objectives. For samples trained without action supervision, we optimize only $\mathcal{L}_{\text{world}}$.

This formulation allows the same WAM backbone to naturally accommodate the different supervision levels in our egocentric data pyramid. During video pre-training, we optimize all samples only with $\mathcal{L}_{\text{world}}$, including both action-valid videos and additional video-only data. During multi-task mid-training, we optimize action-annotated human and robot samples jointly with $\mathcal{L}_{\text{world}}$ and $\mathcal{L}_{\text{action}}$. Post-training follows the same joint objective on downstream robot demonstrations, optionally together with aligned human demonstrations.

B.3 Training Details

We use five training configurations across our experiments: video pre-training, full-scale multi-task mid-training, scaling-study mid-training, real-robot post-training, and RoboDojo post-training. Table 3 summarizes the detailed optimization settings. Unless otherwise specified, all experiments use a per-GPU batch size of 16 with 8 NVIDIA H20 GPUs per node, cosine learning-rate decay, 8-frame future video prediction, and a 32-step action prediction horizon.

For the scaling study, all data-duration and task-diversity variants use the same training configuration and number of optimization steps; only the scale and composition of the egocentric training data are changed. Similarly, all real-robot tasks and robot-data-budget ablations share the same post-training recipe, and all RoboDojo experiments use the same RoboDojo post-training configuration.

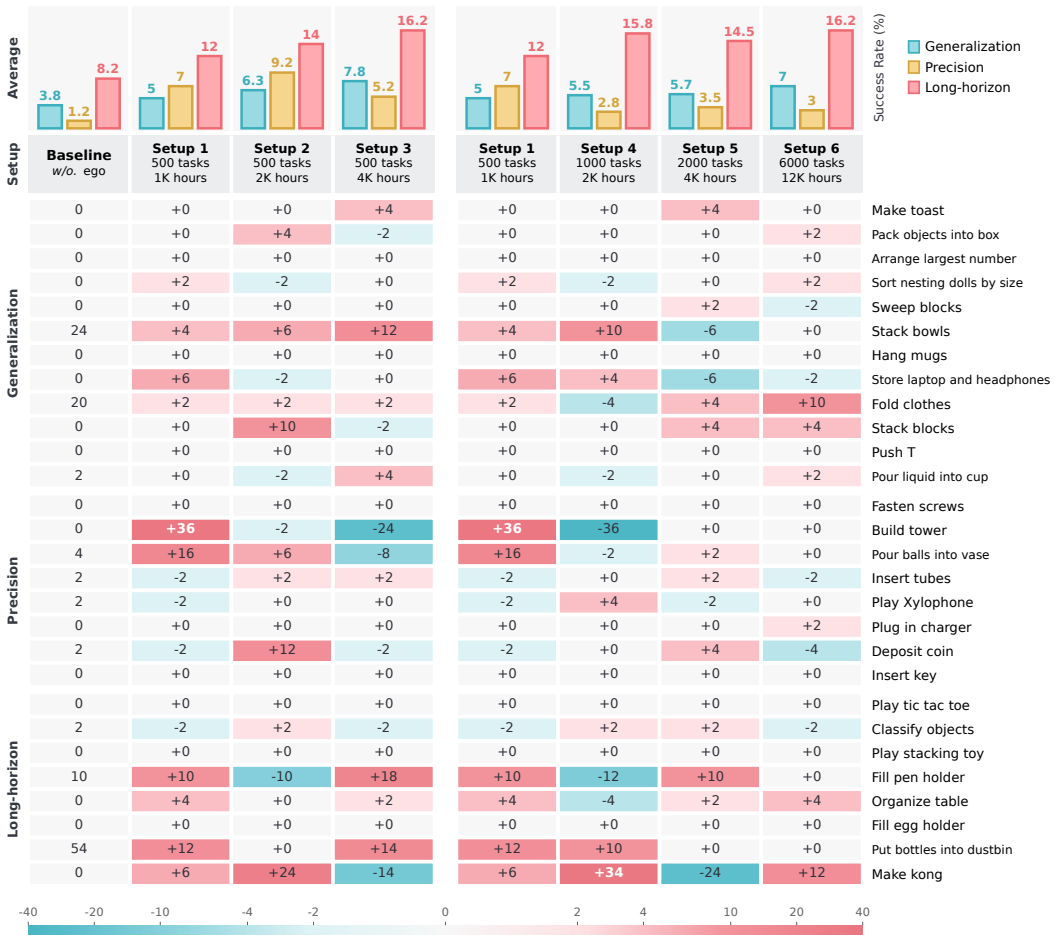


Figure 7: RoboDojo evaluation under data-duration and task-diversity scaling.

C Experiments

C.1 Real-Robot Setup

Robot Platform and Data Collection. We conduct real-world experiments using a bimanual Piper robot equipped with parallel grippers. Robot demonstrations are collected through teleoperation. We record human demonstrations in EgoAlign separately using a monocular head-mounted camera under the corresponding task and environment settings, then convert them into the shared gripper action representation described in Section B.1. Human and robot demonstrations therefore share the same task specification and environment conditions, while retaining their native viewpoints and embodiments.

Evaluation Tasks. Our controlled real-robot experiments focus on two representative manipulation tasks, *Place into the Basket* and *Fold Cloth*, covering rigid-object pick-and-place and deformable-object manipulation, respectively. For each task, we collect 300 in-domain (ID) robot demonstrations. The corresponding EgoAlign set contains 300 ID, 100 object-OOD, and 100 scene-OOD human demonstrations. The object- and scene-OOD conditions are absent from robot training unless otherwise specified, allowing us to evaluate whether human demonstrations can provide information about environment variations not covered by robot data.

Object-OOD evaluation varies the manipulated objects beyond those appearing in the ID robot demonstrations, including changes in object category, instance, appearance, or size. Scene-OOD evaluation changes the surrounding visual environment, including tabletop or tablecloth appearance

and illumination conditions. These two axes capture complementary forms of distribution shift and need not be strictly mutually exclusive.

In addition to these controlled experiments, we evaluate the final model on a broader set of four real-world manipulation tasks: *Fold Towel*, *Tighten a Bottle Cap*, *Stack Plates and Bowls*, and *Heat Food with a Microwave*. These tasks span rigid-object manipulation, deformable-object interaction, and longer-horizon manipulation.

Evaluation Metrics. We evaluate policies through closed-loop execution and report both task success rate and normalized task progress. Success rate measures whether the task is fully completed, while normalized progress provides partial credit for intermediate completion. We use the same evaluation criteria across all compared methods within each task and evaluation setting.

C.2 RoboDojo Setup

We use RoboDojo (Chen et al., 2026) as a standardized closed-loop benchmark for evaluating manipulation policies across a broad range of tasks. We follow the official RoboDojo post-training and evaluation protocol and report both progress score and success rate. All RoboDojo experiments use the same post-training configuration described in Section B.3, allowing comparison of models with different egocentric pre-training and mid-training strategies under a consistent downstream setting.

RoboDojo organizes its evaluation tasks into five categories: *Generalization*, *Precision*, *Long-Horizon*, *Memory*, and *Open*. We report results on all five categories as well as the official average over the complete evaluation suite. Since our model does not explicitly introduce a long-term memory mechanism or a VLM-based module for open-ended instruction following, our analysis in the main paper focuses primarily on *Generalization*, *Precision*, and *Long-Horizon*, while *Memory* and *Open* are retained for completeness.

For the data-scaling study, we compute all reported category averages over the complete RoboDojo evaluation task set. For readability, the task-level visualization in the main paper shows only tasks on which at least one evaluated setting achieves successful rollouts. The complete per-task results, including tasks with zero success across the compared settings, are provided in Figure 7.

C.3 Data Duration and Task-Diversity Scaling

We construct the scaling subsets according to semantic task categories in the egocentric data. The task distribution is naturally imbalanced: a relatively small number of common manipulation behaviors account for a large fraction of the available data, while many less frequent tasks contain substantially fewer demonstrations. We refer to this less frequent portion of the empirical task distribution as the *long tail*.

For duration scaling, we fix the 500 most frequent tasks and progressively increase the amount of data sampled from this fixed task set from 1K to 2K and 4K hours. This increases experience density within a common set of behaviors without expanding task coverage. For task-diversity scaling, we progressively expand the task set from the most frequent categories toward less frequent ones, using 500, 1K, 2K, and 6K tasks. We approximately preserve the average amount of data per task, resulting in 1K, 2K, 4K, and 12K hours of training data, respectively. As the task set grows, we introduce increasingly rare categories from the tail of the empirical task distribution into training.

All variants use the same model initialization, robot-data component, optimization schedule, and RoboDojo post-training protocol. The two scaling sequences therefore emphasize two different ways of increasing egocentric data scale: denser coverage of recurring behaviors and broader task coverage extending toward the long tail.