MiX-World: A Unified World Simulator for Manipulators and Locomoting Humanoids

Anonymous authors
Paper under double-blind review

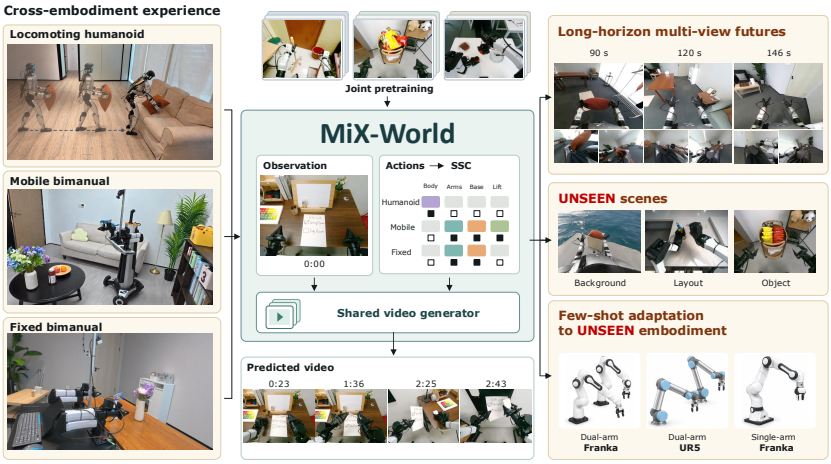


Figure 1: **MiX-World overview.** Joint pretraining across locomoting humanoids and mobile- and fixed-base bimanual manipulators enables a shared action-conditioned model to predict multi-view futures, supporting long-horizon rollouts, unseen-scene generalization, and few-shot adaptation to unseen embodiments. Anonymized project page with videos of all results: `https://anonymous.4open.science/w/mix-world-1BC3/`

Abstract

Action-conditioned world models forecast how visual observations evolve in response to robot actions, yet existing approaches are largely confined to a single robot embodiment. Although robots differ substantially in morphology and control, they interact with the same physical world, suggesting that experience collected across embodiments may provide complementary supervision. However, bringing these experiences together in a single model is challenging. Low-level actions differ across embodiments in both control semantics and action-space parameterization, while trajectories span different physical timescales.

In this paper, we present MiX-World, a unified action-conditioned world simulator for heterogeneous robot embodiments. To our knowledge, it is the first such model to include full-body locomoting humanoids in joint training with mobile-base and fixed-base bimanual manipulators. MiX-World uses Semantically Structured Conditioning (SSC) to represent heterogeneous robot controls, while a shared generator predicts synchronized future observations from multiple views.

To capture both transient interactions and long-horizon evolution, MiX-World introduces Event-Anchored Scheduling (EAS) with Time-Anchored RoPE (TA-RoPE). MiX-World achieves better aggregate visual prediction performance than the evaluated baselines at both short and long horizons. With one demonstration per task and 100 adaptation updates, pretraining also improves full-rollout prediction on three unseen embodiments. Controlled temporal ablations further demonstrate that variable-time training improves both short-timescale event prediction and long-horizon forecasting. MiX-World thus provides a scalable foundation for policy evaluation, interactive data collection, and policy post-training across diverse robot embodiments.

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

1 INTRODUCTION

Recent robot world models have made substantial progress in predicting realistic futures conditioned on robot actions, showing promise as interactive simulators for robot learning (Garcin et al., 2026; Guo et al., 2026; Wei et al., 2026; Wu et al., 2026; Yang et al., 2024; Zheng et al., 2026b; Zhou et al., 2024). These world models largely focus on individual embodiments, although the underlying physics is shared across robots (Gao et al., 2026). Learning from diverse robot experience could therefore improve prediction across embodiments. Recent work has explored cross-embodiment world modeling (He et al., 2026; Wu & Gao, 2026; Huang et al., 2026; Liu & Shorinwa, 2026). However, the benefits of joint training across embodiments as diverse as locomoting humanoids and mobile- and fixed-base manipulators remain underexplored. Realizing this potential requires accommodating differences in action spaces and interaction timescales within a shared world model.

We introduce MiX-World, a unified interactive world simulator spanning these three robot families, with a shared pretrained initialization that transfers to unseen embodiments. The first challenge is to accommodate control interfaces that differ in both dimensionality and physical meaning:

- **Humanoids** combine whole-body motion with hand commands.
- **Mobile-base bimanual manipulators** combine end-effector and gripper commands with planar base-motion and vertical-lift commands.
- **Fixed-base bimanual manipulators** use end-effector and gripper commands.

To this end, MiX-World introduces Semantically Structured Conditioning (SSC) to preserve embodiment-specific control semantics within a shared conditioning interface. Whole-body humanoid actions and end-effector manipulator actions are represented in separate action blocks, while fixed- and mobile-base manipulators share the same bimanual structure. A binary action mask identifies platform-dependent action dimensions, and a shared conditioner injects the resulting action representation into the shared diffusion backbone. This separates what differs across control interfaces while allowing the model to learn visual dynamics jointly across embodiments.

The second challenge is to capture both brief interactions and long-horizon evolution across embodiments with different interaction timescales. Sampling nearby frames captures rapid changes but limits temporal coverage, while sampling frames farther apart covers longer interactions but may miss brief events. We address this trade-off with Event-Anchored Scheduling (EAS), which samples clips over multiple time horizons while retaining frames around detected abrupt action changes. Time-Anchored RoPE (TA-RoPE) encodes elapsed physical time rather than frame index, allowing the model to distinguish intervals of different durations across sampling schedules and dataset recording rates.

MiX-World achieves better aggregate visual prediction than the evaluated embodiment-specific world models at both short and long horizons, and its pretraining transfers to unseen embodiments: on dual-arm Franka and UR5, adapting MiX-World with just 10 demonstrations per task yields better predictions than adapting the baseline without robot pretraining with 100. Beyond prediction, MiX-World provides closed-loop rollouts for policy evaluation, and post-training on its generated trajectories improves real-robot success rates by 15.2–21.2 percentage points across three policies.

Our contributions are:

- To our knowledge, MiX-World is the first cross-embodiment robot world model to include full-body locomoting humanoids in joint training. Specifically, humanoid conditioning combines whole-body motion with hand actions, while the model predicts synchronized moving head and wrist views. SSC lets a single model directly support humanoids, mobile-base manipulators, and fixed-base bimanual manipulators.
- We introduce EAS with TA-RoPE for trajectories with different recording rates and physical durations. Together, they retain brief interaction events across longer temporal horizons and make prediction explicitly aware of elapsed physical time.
- We demonstrate that cross-embodiment pretraining improves adaptation to unseen tasks within the three pretraining families and to three unseen embodiments: dual-arm Franka, dual-arm UR5, and single-arm Franka in LIBERO. These results show that heterogeneous robot experience improves prediction on new embodiments while reducing the amount of target data required for adaptation.

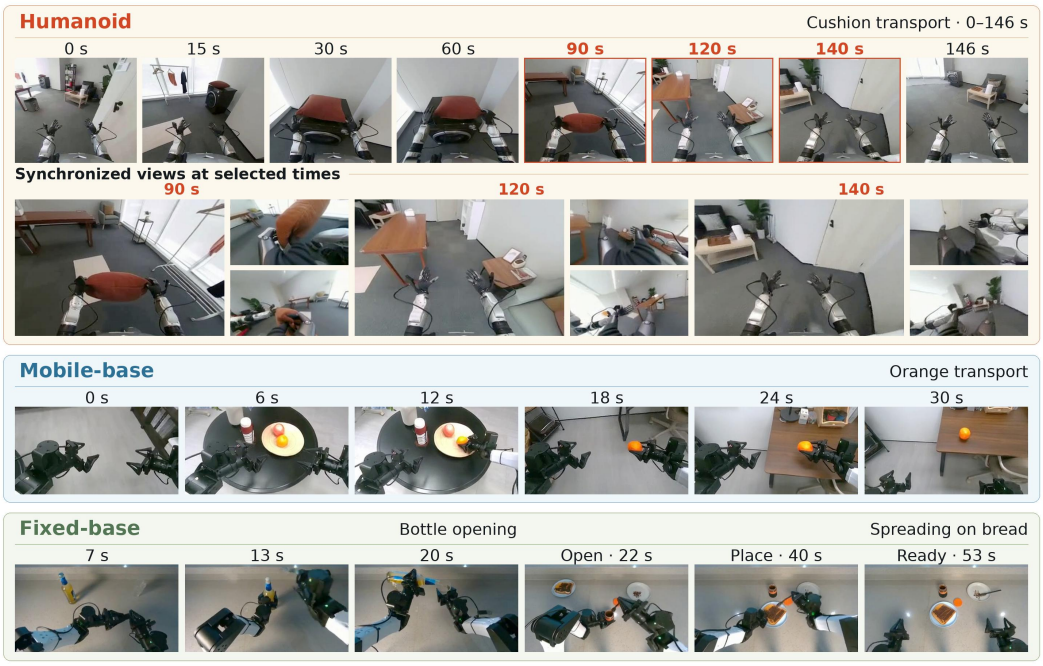


Figure 2: Generated rollouts across the three robot families. Highlighted humanoid timestamps are expanded below into synchronized head and wrist views.

2 METHOD

MiX-World predicts future visual observations conditioned on actions from locomoting humanoids, mobile-base bimanual manipulators, and fixed-base bimanual manipulators. These embodiments differ in their action representations and camera configurations. For locomoting humanoids, whole-body motion changes both the robot's interaction with the environment and its head and wrist viewpoints. The model must therefore predict scene dynamics together with the visual changes caused by camera motion.

2.1 MODEL ARCHITECTURE AND TRAINING

Given a current synchronized multi-view observation, a language instruction, and a timestamped sequence of actions, MiX-World predicts future multi-view observations. As shown in Figure 3, MiX-World builds on a pretrained Wan latent video generator (Wan et al., 2025). The synchronized multi-view videos are spatially combined into a single video and encoded by the VAE. The diffusion transformer predicts the future video latents conditioned on the initial observation, language instruction, and SSC inputs (Section 2.2). Physical time enters through TA-RoPE and intra-group timing conditioning (Section 2.3). The VAE decodes the predicted latents into multi-view canvases, so all views are generated jointly. Backbone and initialization details are provided in Appendix C, and camera configurations in Appendix G.

We jointly train MiX-World on in-house and public robot trajectories spanning humanoids, mobile-base bimanual manipulators, and fixed-base bimanual manipulators. The video backbone, SSC conditioner, and physical-time conditioner are optimized using the flow-matching objective, while the VAE and pretrained image and text encoders remain frozen.

2.2 SEMANTICALLY STRUCTURED CONDITIONING

SSC maps heterogeneous actions into a common vector. Action fields with shared physical meaning use the same coordinates, while fields with different meanings occupy separate blocks. A binary mask identifies the active coordinates for each embodiment.

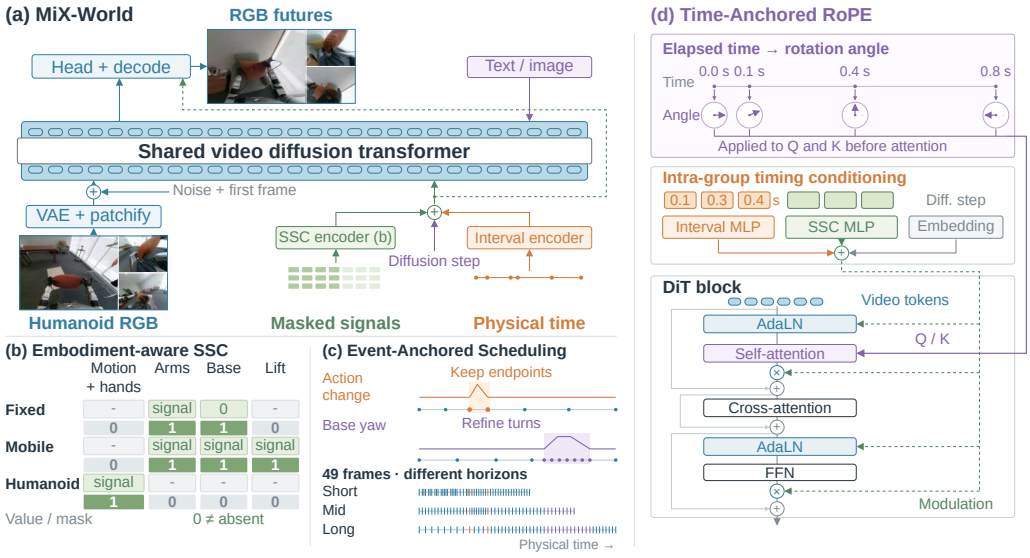


Figure 3: MiX-World architecture and conditioning design. (a) Joint multi-view prediction. (b) SSC with shared and platform-specific masked action blocks. (c) EAS preserves brief transitions across temporal scales. (d) TA-RoPE and intra-group timing conditioning encode physical time.

Union conditioning representation. For G1, a dedicated block contains a discrete whole-body motion token from the SONIC encoder (Luo et al., 2026) and two hand commands. Manipulator actions occupy a separate block, with shared end-effector, gripper, and base fields for fixed- and mobile-base embodiments and an additional lift field for mobile-base embodiments.

Let $s_t^{(e)}$ denote the action representation of embodiment e at time t. Embodiment-specific preprocessing Φ_e maps it to a common vector $u_t^{(e)} \in \mathbb{R}^D$ and a validity mask $m_t^{(e)} \in \{0,1\}^D$. The conditioning module receives

$$r_t^{(e)} = [\, u_t^{(e)} \odot m_t^{(e)}; m_t^{(e)} \,], \qquad (u_t^{(e)}, m_t^{(e)}) = \Phi_e(s_t^{(e)}). \tag{1}$$

The mask distinguishes valid zero values from unused fields. For example, fixed-base samples have zero base values with an active mask; the same fields contain base commands for mobile-base samples, including zeros when stationary.

Shared conditioning module. We group the masked action representations to match the VAE's temporal compression. A shared conditioner f_s processes each group and produces a residual δh_i^s for the diffusion-time embedding and a transformer modulation δM_i^s for latent frame i. The same conditioner processes actions from all embodiments. Appendix C.1 provides implementation details, and Table 4 compares SSC with zero-padded joint training.

2.3 Event-Anchored Scheduling with Time-Anchored RoPE

EAS samples trajectories at multiple temporal scales while retaining frames around abrupt action changes. TA-RoPE encodes the physical times of the sampled frames.

Event-Anchored Scheduling (EAS). We score changes between adjacent actions within semantic groups and detect abrupt changes using a threshold for each group. Frames on both sides of each detected change are marked for inclusion. We then sample at multiple temporal scales, shortening any proposed gap that would skip a marked frame. Overlapping clips are added until every detected transition is covered by at least one clip. The same frame indices select all synchronized camera views and their corresponding actions. Detection thresholds, sampling settings, and an additional base-rotation constraint are detailed in Appendix C.2.

Time-Anchored Rotary Position Embedding (TA-RoPE). Nonuniform sampling decouples a frame's sequence index from its physical time. Let k_j be the raw-frame index of sampled frame j and

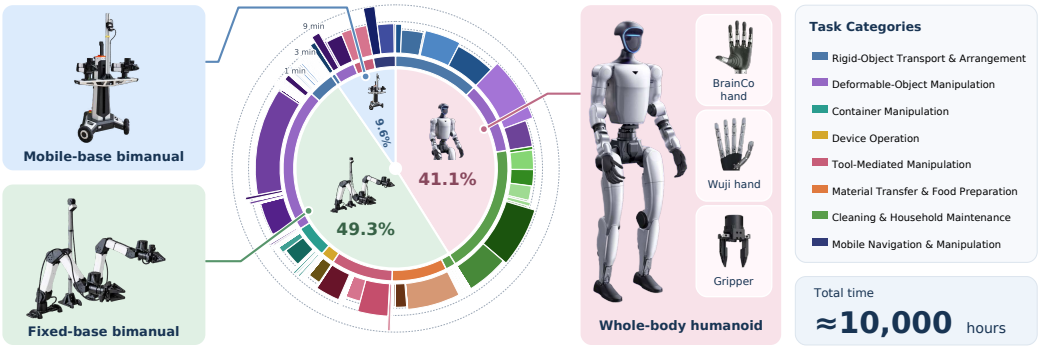


Figure 4: Composition of the pretraining corpus across embodiments and task categories. Wedge angle and radius encode interaction hours and mean trajectory duration, respectively.

f_e the recording rate of dataset e. Its elapsed time τ_j determines the RoPE position p_i of latent-frame anchor i:

$$\tau_j = \frac{k_j - k_0}{f_e}, \qquad p_i = \frac{\tau_{qi}}{q\Delta\tau_{\text{ref}}}, \tag{2}$$

where q is the temporal grouping factor and $\Delta\tau_{\text{ref}}$ is a reference frame interval. Uniform sampling at this interval gives $p_i = i$, recovering the standard index-based positions. With nonuniform sampling, RoPE uses the continuous positions p_i with the same rotary frequencies, so its phase advances with elapsed physical time.

Intra-group timing conditioning. An anchor position specifies when a latent frame occurs, but does not describe the individual time gaps within its action group. We therefore also encode the intervals $\Delta\tau_j = \tau_{j+1} - \tau_j$ within each group using a physical-time conditioner. Its outputs are added to the SSC residual δh_i^s and modulation δM_i^s. These inputs describe elapsed trajectory time, distinct from the diffusion timestep used during generation. Feature definitions and padding rules are provided in Appendix C.2.

3 Experiments

We evaluate **MiX-World** as a unified world model for heterogeneous robots, focusing on prediction quality, transfer and generalization, and its utility for policy learning. To assess these capabilities and the contributions of action conditioning and temporal modeling, we organize our experiments around five research questions:

- (RQ1) Can a single jointly pretrained **MiX-World** checkpoint follow each embodiment's action while maintaining strong visual prediction quality over both short and long horizons?
- (RQ2) Does cross-embodiment pretraining enable faster, more data-efficient adaptation to new tasks and unseen embodiments?
- (RQ3) Does EAS with TA-RoPE improve prediction during concentrated motion and interaction events and align predicted motion progression with elapsed physical time?
- (RQ4) Can **MiX-World** generalize conditional prediction to unseen scenes?
- (RQ5) Does **MiX-World** support policy evaluation and effective policy post-training?

Appendix A summarizes the experimental evidence. We first describe the training corpus and comparison protocols.

3.1 Experimental Setup

We train **MiX-World** on the three robot families introduced in Section 1, spanning a whole-body humanoid, a fixed-base bimanual manipulator, and a mobile-base bimanual manipulator. Most of the corpus is collected in-house through teleoperation on these robots; it is complemented by the public

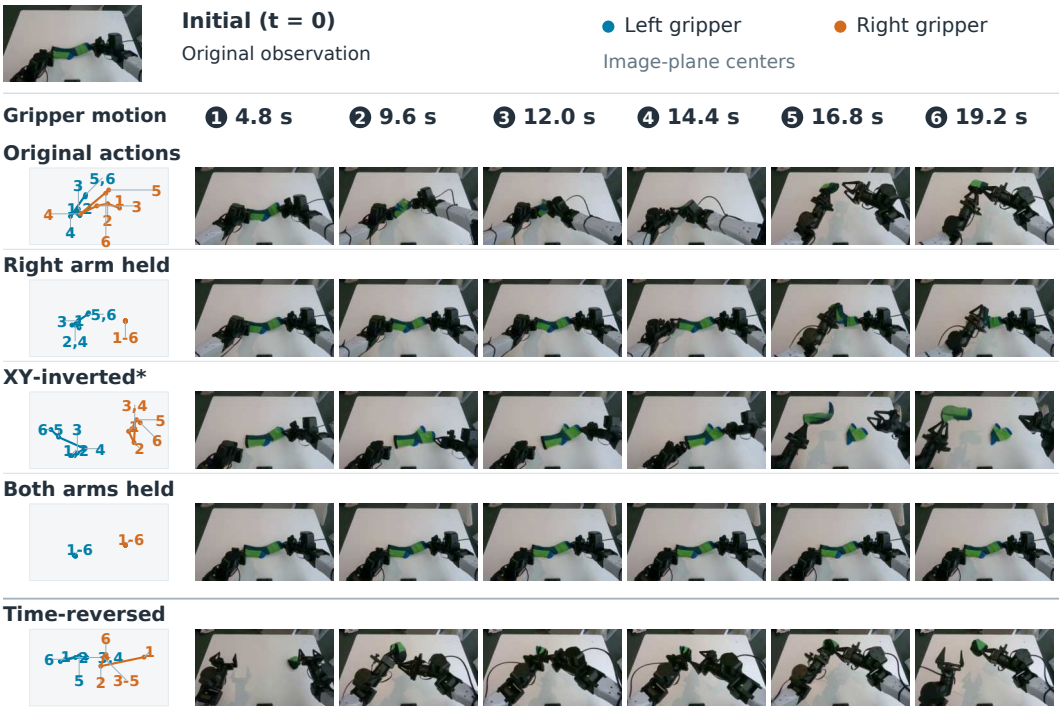


Figure 5: Action-conditioned fixed-base bimanual rollouts under original and modified action sequences. The first four rows start from the same observation, while the time-reversed rollout starts from the original rollout's generated endpoint. The resulting differences illustrate qualitative sensitivity to action conditioning.

humanoid datasets HIW-500 (BitRobot Foundation, 2026) and Unitree WBT (Unitree Robotics, 2026) and by two public bimanual corpora, ABC-130K (Allshire et al., 2026) and MolmoAct2 (Fang et al., 2026). In total, the corpus contains roughly 10,000 hours of robot interaction. Humanoid locomotion is encoded as 64-dimensional discrete SONIC motion tokens (Luo et al., 2026) together with two hand commands, synchronized with one head-camera and two wrist-camera streams. Figure 4 summarizes the corpus composition across embodiments and task categories. The 26 subtask-level interaction categories behind this overview are illustrated in Figure 14 in the appendix.

Controlled comparisons. Table 1 compares one jointly trained **MiX-World** checkpoint with embodiment-specific baselines, matching training-data amounts and epoch counts for each task (Appendix C.3). For target adaptation, both initializations use the pretrained Wan video backbone. The baseline omits cross-embodiment robot pretraining; figures and tables use the short label "w/o robot pretrain". Paired runs use identical target examples and matched optimizer updates. On unseen embodiments, both also use the same action-conditioned architecture (Appendix E.1). We evaluate SSC against zero-padded joint training on the same multi-embodiment corpus (Table 4) and compare EAS with TA-RoPE against fixed-rate sampling (Table 2).

3.2 Action Controllability and Prediction Quality

We evaluate MiX-World's response to action interventions in a bimanual manipulation task (Figure 5). We modify arm commands and also test a time-reversed rollout starting from the original prediction's final frame. Changing arm commands produces different robot motions and changes in the positions and shapes of deformable objects, while holding both arms keeps the scene nearly static. These results provide qualitative evidence that action inputs influence both robot motion and the predicted state of the manipulated object. Additional mobile-base examples appear in Appendix D.1.

Table 1 compares visual prediction quality between **MiX-World** and Ctrl-World (Guo et al., 2026), Motus (Bi et al., 2026), and DreamDojo (Gao et al., 2026), using predictions from all three camera views. At both short and long horizons, MiX-World achieves the best aggregate scores across all

Table 1: Visual prediction quality at short (4.8 s) and autoregressive long (96 s) horizons. All methods receive the same initial observations and actions. LPIPS, SSIM, and PSNR are macro-averaged across three embodiments using three-view composites. MiX-World's per-embodiment PSNR, SSIM, and LPIPS results appear in the SSC rows of Table 4.

	Short horizon			Long horizon		
Method	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
Ctrl-World (Guo et al., 2026)	0.204	0.772	21.695	0.403	0.580	14.725
Motus (Bi et al., 2026)	0.228	0.734	17.538	0.312	0.621	16.474
DreamDojo (Gao et al., 2026)	0.192	0.807	22.761	0.342	0.634	15.923
MiX-World	**0.178**	**0.829**	**24.431**	**0.290**	**0.666**	**18.259**

reported visual prediction metrics. These results demonstrate improved overall prediction quality under the recorded robot actions. Qualitative comparisons with ground truth are shown in Figure 10 and Appendix C.4.

Effect of SSC. Table 4 reports the per-embodiment results of MiX-World alongside a zero-padding baseline trained jointly on the same corpus. SSC improves PSNR, SSIM, and LPIPS for every robot family at both horizons. Averaged equally across the three families, the PSNR gain is 3.96 dB at 4.8 s and 5.87 dB at 96 s. Humanoids show the largest PSNR gains, and each family benefits more at the long horizon. These results show the value of structured action conditioning for prediction across heterogeneous robot bodies.

3.3 Cross-Embodiment Pretraining Improves Target Adaptation

Figure 6 compares target-task fine-tuning with and without cross-embodiment robot pretraining within the three pretraining families and on three unseen embodiments, under the controls in Section 3.1.

Seen robot families, unseen target tasks. MiX-World pretraining improves adaptation to unseen tasks. Across the fixed-base, mobile-base, and humanoid evaluations, MiX-World initialization achieves higher SSIM than the baseline without robot pretraining at all evaluated checkpoints (Figure 6(a)). At representative checkpoints, MiX-World initialization also reduces prediction errors in robot-arm and manipulated-object regions compared with target adaptation without robot pretraining (Figure 6(b); Table 10). Object MSE decreases by 70.98% for fixed-base manipulation and 57.62% for mobile manipulation. These object evaluations cover the interaction interval from first contact onward. Together with the Arm MSE results, these paired comparisons show that the adaptation gains extend to visual predictions of the robot and the objects it manipulates.

Unseen embodiments. MiX-World pretraining improves adaptation to unseen embodiments, including dual-arm Franka and UR5 in RoboTwin 2.0 (Chen et al., 2025) and single-arm Franka in LIBERO (Liu et al., 2023). It achieves higher full-rollout PSNR and SSIM across all evaluated settings (Figure 6(c); Table 6); metrics report mean $\pm$ sample standard deviation over multiple inference seeds. On both dual-arm robots, 10-demo MiX-World outperforms the 100-demo baseline without robot pretraining even when that baseline receives up to five times as many updates. Since these embodiments are absent from pretraining, the results support transfer of predictive knowledge to new robot bodies. The tenfold difference in demonstration budgets further highlights the value of cross-embodiment pretraining when target data are limited.

3.4 Event Prediction and Physical-Time Alignment

We compare EAS with TA-RoPE against fixed-rate sampling to assess its effect on predicting robot motion and interaction. The combined design improves prediction quality on both the base-motion and bimanual-interaction subsets, as well as the full evaluation set, and outperforms either component alone (Table 2). With the number of observations per training sample held fixed, these gains indicate that the temporal design helps the model use the available training observations more effectively.

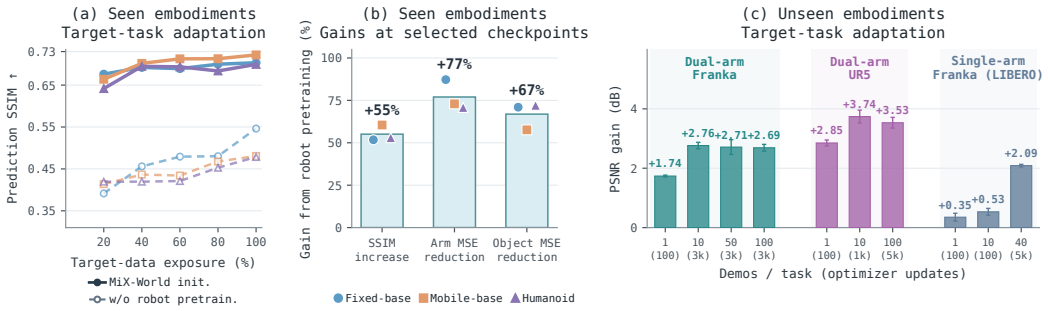


Figure 6: Pretraining improves adaptation on seen and unseen embodiments. (a) SSIM on unseen tasks within pretrained robot families. (b) Gains from cross-embodiment robot pretraining at representative checkpoints. (c) Seed-averaged full-rollout test PSNR gains (dB) from this pretraining on unseen embodiments; error bars show ± s.d. over multiple inference seeds.

The base-rotation example further shows how this improvement appears in predicted motion (Figure 7). Compared with fixed-rate sampling, EAS with TA-RoPE predicts positions and orientations closer to ground truth at corresponding physical times, providing qualitative evidence of more accurate motion progression.

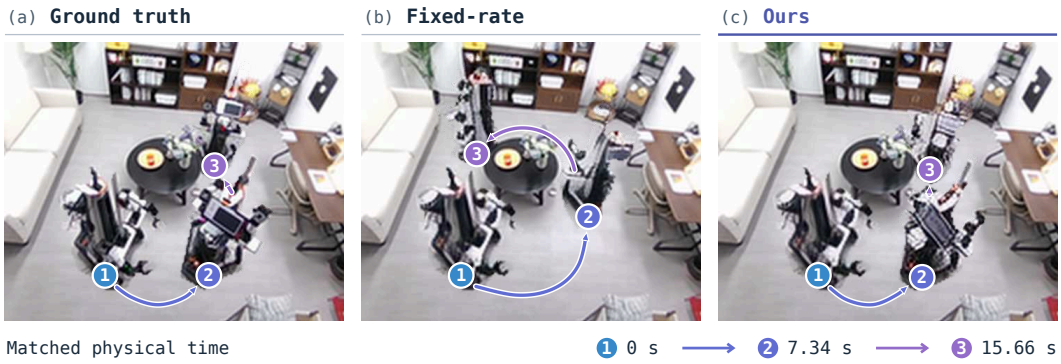


Figure 7: Motion prediction at matched physical times during base rotation. EAS with TA-RoPE follows the ground-truth motion more closely than fixed-rate sampling. Both models predict future ceiling views jointly with the robot-mounted views. Original four-view frames appear in Appendix F.1.

Table 2: Prediction quality on the mobile manipulator. We compare EAS with TA-RoPE against fixed-rate sampling and single-component variants on the full evaluation set and two subsets focused on base motion and bimanual interaction. Prediction quality improves across all three settings, with the largest gains during base motion.

	Overall			Mobile-Base Motion			Bimanual Manipulation		
Temporal training	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fixed-rate sampling	15.525	0.5773	0.2413	15.243	0.6019	0.2298	16.113	0.6090	0.2232
w/o EAS	18.273	0.6479	0.1543	18.512	0.6655	0.1316	18.534	0.6699	0.1573
w/o TA-RoPE	18.353	0.6550	0.1506	18.343	0.6651	0.1431	18.756	0.6823	0.1479
EAS + TA-RoPE (Ours)	**18.631**	**0.6628**	**0.1415**	**18.682**	**0.6707**	**0.1297**	**18.939**	**0.6890**	**0.1443**

3.5 Generalization to Constructed Unseen Scenes

Figure 8 qualitatively evaluates generalization to unseen scenes. Starting from one source episode, we keep the recorded fixed-arm action, prompt, checkpoint, seed, and sampling setup fixed, and alter

one visual factor at a time: the background scene, the counter layout, or the identity of a familiar object. Additional examples are provided in Appendix D.2.

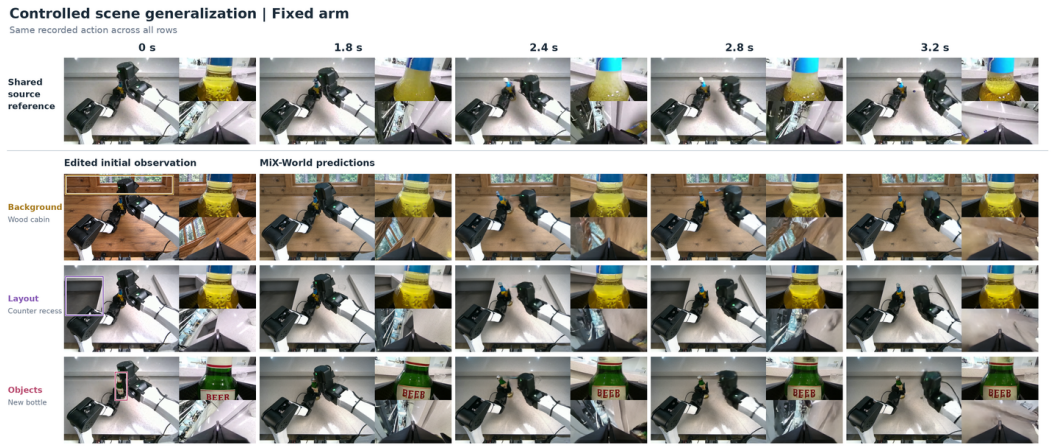


Figure 8: Unseen-scene prediction under shared fixed-arm actions after background, layout, or object edits. The top row is the unedited source reference, not counterfactual ground truth. Edited rows show initial observations ($t = 0$) and predicted head/wrist views at 1.8–3.2 seconds.

3.6 MiX-World as a Policy Evaluator and Data Generator

Table 3 evaluates MiX-World as both a policy evaluator and a data generator. All entries are task success rates. The two base-policy columns evaluate the same checkpoint before post-training, either on the real robot or through closed-loop MiX-World rollouts. The final column reports real-robot success after further training that policy on MiX-World-generated trajectories (protocol in Appendix C.6). As a policy evaluator, MiX-World reproduces the real-robot ranking of all three policies, supporting its use for policy ranking. As a data generator, it provides trajectories whose use in post-training improves real-robot success across all three policies.

Table 3: Task success rates (%) for policy evaluation and post-training. MiX-World preserves the real-robot policy ranking, and post-training on its generated trajectories improves real-robot success across all three policies.

Policy	Base (Real)	Base (MiX-World)	Post-trained (Real)
π_0 (Black et al., 2024)	21.2%	27.3%	42.4%
$\pi_{0.5}$ (Physical Intelligence, 2025)	57.6%	62.1%	78.8%
Diffusion Policy (Chi et al., 2023)	12.1%	9.1%	27.3%

4 Conclusion

MiX-World unifies fixed-base manipulation, mobile manipulation, and humanoid locomotion in a shared multi-view world simulator through semantically structured conditioning and physical-time-aware training. A single jointly pretrained checkpoint improves aggregate visual prediction over embodiment-specific baselines at short and long horizons and supports data-efficient adaptation to three unseen embodiments. Post-training with generated trajectories and original data improves real-world success rates by 15.2–21.2 percentage points across three policies.

AI USAGE STATEMENT

During the preparation of this paper, we used generative AI tools to assist with writing and language editing, including improving wording, grammar, and clarity, as well as refining text organization. All AI-assisted revisions incorporated into the paper were reviewed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the final content, including its scientific claims, experimental results, and references.

REFERENCES

1X World Model Team. 1X World Model: Evaluating bits, not atoms. https://www.1x.tech/1x-world-model.pdf, 2025.

Muhammad Qasim Ali, Aditya Sridhar, Shahbuland Matiana, Alex Wong, and Mohammad Al-Sharman. Humanoid world models: Open world foundation models for humanoid robotics. *arXiv preprint arXiv:2506.01182*, 2025.

Arthur Allshire, Himanshu Gaurav Singh, Ritvik Singh, Adam Rashid, Hongsuk Choi, David McAllister, Justin Yu, Yiyuan Chen, Huang Huang, Pieter Abbeel, et al. Scalable behavior cloning with open data, training, and evaluation. *arXiv preprint arXiv:2606.27375*, 2026. ABC-130K dataset: https://huggingface.co/datasets/XDOF/ABC-130k.

Anurag Bagchi, Zhipeng Bao, Homanga Bharadhwaj, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Walk through paintings: Egocentric world models from internet priors. *arXiv preprint arXiv:2601.15284*, 2026.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

BitRobot Foundation. HIW-500: Humanoids in-the-wild dataset. https://huggingface.co/datasets/BitRobot/HIW-500, 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Shuang Zeng, Junjin Xiao, Xinyuan Chang, Mu Xu, et al. ABot-PhysWorld: Interactive world foundation model for robotic manipulation with physics alignment. *arXiv preprint arXiv:2603.23376*, 2026.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

Haoquan Fang, Jiafei Duan, Donovan Clay, Sam Wang, Shuo Liu, Weikai Huang, Xiang Fan, Wei-Chuan Tsai, Shirui Chen, Yi Ru Wang, et al. MolmoAct2: Action reasoning models for real-world deployment. *arXiv preprint arXiv:2605.02881*, 2026.

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

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3D state. In *International Conference on Machine Learning*, 2026.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, 2026.

Zihao He, Bo Ai, Tongzhou Mu, Yulin Liu, Weikang Wan, Jiawei Fu, Yilun Du, Henrik I. Christensen, and Hao Su. Scaling cross-embodiment world models for dexterous manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2026.

Ze Huang, Jiahui Zhang, Hairuo Liu, Chenxi Zhang, Ran Cheng, and Li Zhang. Learning transferable dynamics priors from action to world modeling. In *European Conference on Computer Vision*, 2026.

Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025a.

Zhennan Jiang, Kai Liu, Yuxin Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Chao Yu, Haoran Li, and Dongbin Zhao. World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. *arXiv preprint arXiv:2509.19080*, 2025b.

Zhennan Jiang, Shangqing Zhou, Yutong Jiang, Zefang Huang, Mingjie Wei, Yuhui Chen, Tianxing Zhou, Zhen Guo, Hao Lin, Quanlu Zhang, et al. Wovr: World models as reliable simulators for post-training vla policies with rl. *arXiv preprint arXiv:2602.13977*, 2026.

Byungjun Kim, Taeksoo Kim, Hyunsoo Cha, and Hanbyul Joo. Robot-factored world models via robot rendering. *arXiv preprint arXiv:2607.22535*, 2026.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. WorldEval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

Yaxuan Li, Zhongyi Zhou, Yefei Chen, Yanjiang Guo, Jiaming Liu, Shanghang Zhang, Jianyu Chen, and Yichen Zhu. Hi-WM: Human-in-the-world-model for scalable robot post-training. *arXiv preprint arXiv:2604.21741*, 2026a.

Yaxuan Li, Zhongyi Zhou, Yefei Chen, Yaokai Xue, and Yichen Zhu. dWorldEval: Scalable robotic policy evaluation via discrete diffusion world model. *arXiv preprint arXiv:2604.22152*, 2026b.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. URL https://arxiv.org/abs/2306.03310.

Kechen Liu and Ola Shorinwa. CLAP: Cross-embodiment video world models are zero-shot physical simulators. *arXiv preprint arXiv:2608.27406*, 2026.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. doi: 10.1126/scirobotics.aed4592. URL https://www.science.org/doi/10.1126/scirobotics.aed4592.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations*, 2023.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. MotuBrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, 2022.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, et al. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation*, 2024.

Physical Intelligence. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Ansh Kumar Sharma, Yixiang Sun, Ninghao Lu, Yunzhe Zhang, Jiarao Liu, and Sherry Yang. World-gymnast: Training robots with reinforcement learning in a world model. *arXiv preprint arXiv:2602.02454*, 2026.

Haofeng Sun, Jiangbo Pei, Fei Kang, Zexiang Liu, Yaokun Li, Boyi Jiang, Hua Xue, Cindy Zhou, Wei Li, Yichen Wei, et al. Riemann-1.0: An embodied world action model for physical AI. *arXiv preprint arXiv:2608.27033*, 2026.

Gemini Robotics Team, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Fangchen Liu, Anirudha Majumdar, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

Unitree Robotics. UnifoLM-WBT: Unitree G1 whole-body teleoperation dataset. `https://huggingface.co/collections/unitreerobotics/unifolm-wbt-dataset`, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL `https://arxiv.org/abs/2503.20314`.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning*, 2023.

Yixuan Wang et al. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026.

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, and Qi Fan. Geometry-aware implicit memory for video world models. *arXiv preprint arXiv:2606.02436*, 2026.

Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. MultiWorld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564*, 2026.

Zhuoyuan Wu and Jun Gao. OSCAR: Omni-embodiment action-conditioned world model for robotics. *arXiv preprint arXiv:2606.04463*, 2026.

Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. Kinema4D: Kinematic 4D world modeling for spatiotemporal embodied simulation. *arXiv preprint arXiv:2603.16669*, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024.

Jia Zheng, Teli Ma, Yudong Fan, Zifan Wang, Shuo Yang, and Junwei Liang. MotionWAM: Towards foundation world action models for real-time humanoid loco-manipulation. *arXiv preprint arXiv:2606.09215*, 2026a.

Zirui Zheng, Jiaqian Yu, Xiongfeng Peng, Jun Shi, Mingyi Li, Chao Zhang, Weiming Li, Dong Wang, Huchuan Lu, and Xu Jia. Mem-World: Memory-augmented action-conditioned world models for persistent robot manipulation. *arXiv preprint arXiv:2606.18960*, 2026b.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

A Overview of Experimental Evidence

We evaluate MiX-World through complementary prediction, adaptation, intervention, and policy-learning experiments. First, a single jointly trained checkpoint supports action-conditioned, multi-view prediction across fixed-base manipulators, mobile manipulators, and humanoid robots. Table 1 compares this shared model with embodiment-specific baselines over 4.8-second and 96-second horizons, using identical initial observations and action sequences; MiX-World achieves the best aggregate results on all three reported metrics at both horizons. Same-architecture comparisons further show that joint multi-embodiment training improves PSNR, SSIM, and LPIPS over single-embodiment training for all three robot families at both horizons (Table 5). Second, adaptation experiments cover unseen target tasks within the three pretrained robot families and three previously unseen embodiments. Within the pretrained families, MiX-World initialization achieves higher SSIM than the baseline without robot pretraining at all evaluated checkpoints (Figure 6(a)). For unseen embodiments, we reuse the same pretrained initialization. Paired runs use the same architecture, target demonstrations, optimizer updates, and test trajectories. MiX-World improves both PSNR and SSIM in all eleven matched fine-tuning settings; on dual-arm Franka and UR5, ten demonstrations per task outperform the baseline without robot pretraining trained on one hundred demonstrations, even when that baseline receives up to five times as many updates (Table 6). These unseen-embodiment results report mean and sample standard deviation over multiple inference seeds. Tables 8 and 9 further report results by task and camera view.

We also examine action conditioning, temporal modeling, and responses to action interventions and controlled first-frame scene edits. SSC improves PSNR, SSIM, and LPIPS over zero-padded joint training for all three robot families at both horizons (Table 4). The comparison of EAS with TA-RoPE against fixed-rate sampling and single-component variants covers 128 mobile-manipulator trajectories, including subsets focused on base motion and bimanual interaction (Table 2). Visual interventions span all three robot families and expose how predictions respond to changed actions or scene conditions. Finally, closed-loop evaluation covers three policies, and policies post-trained with 500 generated trajectories together with original data achieve real-robot success gains of 15.2–21.2 percentage points over their base checkpoints (Table 3). Appendices C–E provide evaluation protocols, additional intervention examples, adaptation controls, and detailed quantitative breakdowns.

B Related Work

Action-conditioned robot world models. Generative interactive environments such as Genie show that video models can simulate plausible futures under provided controls (Bruce et al., 2024). Building on pretrained video generation backbones, recent robot world models adapt this generative paradigm to predict future observations conditioned on robot actions. Ctrl-World, Genie Envisioner, and OSCAR improve action following with learned or rendered motion conditions (Guo et al., 2026; Jiang et al., 2025a; Liao et al., 2025; Xu et al., 2026; Wu & Gao, 2026; Kim et al., 2026). Other methods focus on consistent prediction across views and over longer horizons. MultiWorld jointly predicts multiple views, while Mem-World introduces memory for long rollouts (Wu et al., 2026; Huang et al., 2026; Zheng et al., 2026b; Wei et al., 2026; Garcin et al., 2026). Accurate physical interaction additionally depends on learned priors for contact and object response. Their quality is visible in whether objects begin moving at contact, respond to pushes, and remain coupled to the robot for the appropriate duration (Chen et al., 2026; Kim et al., 2026; Wang et al., 2026). Such errors are especially consequential in policy evaluation and post-training, where an incorrect interaction outcome can misjudge a policy or provide invalid training experience (Li et al., 2025; 2026b; Quevedo et al., 2025; Team et al., 2025; Jiang et al., 2025b; Sharma et al., 2026; Jiang et al., 2026; Li et al., 2026a). MiX-World studies how physical interaction priors learned across robot embodiments improve prediction under platform-native conditioning.

Humanoid world models. Humanoid prediction has recently emerged as a demanding setting for video world models. 1XWM forecasts full-body humanoid interaction from low-level action sequences and uses the resulting rollouts for policy evaluation (1X World Model Team, 2025). HWM studies action-conditioned egocentric prediction on humanoid demonstrations (Ali et al., 2025), while EgoWM injects motor commands into pretrained video models across navigation platforms and a 25-DoF humanoid (Bagchi et al., 2026). MotionWAM further couples a video dynamics prior to unified whole-body motion generation for real-time humanoid loco-manipulation (Zheng et al.,

2026a). These studies establish the value of full-body humanoid prediction, with their reported humanoid models trained or adapted around a target platform and control interface. MiX-World places locomoting humanoids inside a shared operating model: one jointly pretrained checkpoint directly supports humanoid, mobile-base bimanual, and fixed-base bimanual robots, while retaining optional target specialization.

Cross-embodiment learning. Cross-embodiment learning has benefited robot policies, representations, imitation learning, and world-action models (Open X-Embodiment Collaboration et al., 2024; Octo Model Team et al., 2024; Nair et al., 2022; Ma et al., 2023; Wang et al., 2023; Bi et al., 2026; MotuBrain Team et al., 2026; Sun et al., 2026). CrossWM, OSCAR, and A2World share data through particle, skeleton, and end-effector action representations (He et al., 2026; Gao et al., 2026; Wu & Gao, 2026; Huang et al., 2026). CLAP combines end-effector, language, and latent actions for cross-embodiment pretraining and adapts the learned prior to bimanual and humanoid platforms (Liu & Shorinwa, 2026). These methods make heterogeneous experience transferable by projecting control into shared particle, skeleton, end-effector, language, or latent abstractions. This makes data easier to pool, while the chosen abstraction determines which embodiment-specific motion remains visible to the predictor. Their reported evaluations primarily center on manipulation-compatible interfaces or target-specific models. MiX-World jointly models locomoting humanoids and mobile- and fixed-base bimanual manipulators through semantically structured conditioning and physical-time-aware training. Beyond supporting heterogeneous control interfaces in one checkpoint, we demonstrate that experience from other embodiments improves prediction on a target robot in controlled transfer experiments. Our study thus examines both how heterogeneous robotic experience can be modeled jointly and when it benefits target prediction.

C Implementation and Evaluation Details

Backbone and conditioning initialization. MiX-World is initialized from the `Wan2.1-Fun-V1.1-14B-InP` latent video generator. We add an SSC conditioner and a local physical-time conditioner to its diffusion transformer. Both conditioners use zero-initialized output projections, so their added contributions are zero at initialization. TA-RoPE separately encodes elapsed physical time through the temporal rotary positions.

Data splits and optimization. For each task, we split the data into training and test sets at the episode level. We train with AdamW using a learning rate of 5×10^{-5} and a batch size of 32.

C.1 SSC implementation details

Action dimensions. The G1 block contains a 64-dimensional discrete whole-body motion token produced by the SONIC encoder (Luo et al., 2026), together with two hand commands, giving 66 dimensions. SSC combines this block with a separate 18-dimensional manipulator block in an 84-dimensional action space. The manipulator fields are allocated as follows:

Action fields	Dimensions
Two end-effector poses and gripper commands	14
Planar base motion	3
Lift position	1

Fixed- and mobile-base samples both activate the end-effector, gripper, and base fields. Fixed-base samples set the base values to zero; mobile-base samples provide base commands and additionally activate the lift field. Coordinates not used by an embodiment have an inactive mask.

Temporal grouping and conditioning. The video VAE maps 49 observations to 13 latent frames. We concatenate four consecutive masked action representations for each latent frame, repeating the terminal representation to complete the final group:

$$c_i^s = \mathrm{LN}\Big(\mathrm{Concat}_{j=0}^{3}\, r_{4i+j}^{(e)}\Big), \qquad (\delta h_i^s, \delta M_i^s) = f_s(c_i^s), \tag{3}$$

where $i = 0, \ldots, 12$ and $r_t^{(e)} = r_{48}^{(e)}$ for $t > 48$. The two MLP heads of f_s produce a residual for the diffusion-time embedding and a transformer modulation, respectively.

SSC ablation. Figure 9 contrasts direct action mixing with SSC. We compare SSC with a direct action-mixing baseline (zero-padding) trained jointly on humanoid, fixed-base bimanual, and mobile-base bimanual data. The baseline uses a shared 66-dimensional action space. Humanoid actions comprise a 64-dimensional SONIC motion token and two hand commands. Bimanual actions comprise 14 end-effector and gripper coordinates, with mobile-base robots additionally providing three planar base-velocity components and one lift position. Each representation is placed at the beginning of the shared vector, and unused coordinates are zero-padded, so coordinates with different physical meanings overlap across embodiments. SSC allocates separate 66-dimensional humanoid and 18-dimensional manipulator blocks in an 84-dimensional space, preserving shared manipulator semantics while separating physically distinct controls.

Both models normalize actions using frozen training-set 1st- and 99th-percentile statistics and clip them to $[-1, 1]$. Fixed-base samples encode zero physical base velocity with an active mask; the unsupported lift coordinate is masked out. Each masked action vector is concatenated with its binary validity mask, grouped over four consecutive timesteps, and passed through a shared conditioner that produces diffusion-time embedding residuals and transformer modulation signals. Initialization, training data, optimizer, training budget, and EAS with TA-RoPE settings are identical. We evaluate both models using the short- and long-horizon protocol in Appendix C.3. The SSC rows in Table 4 report the per-embodiment results of the MiX-World checkpoint in Table 1.

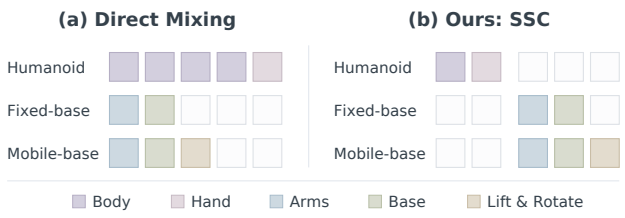


Figure 9: Direct action mixing and SSC. SSC separates humanoid body and hand controls from shared manipulator fields. Colored blocks denote active action fields, and empty blocks denote unused coordinates. Lift and rotation controls are grouped for mobile-base robots.

SSC improves all three metrics for every robot family at both horizons. For fixed-base, mobile-base, and humanoid robots, respectively, PSNR increases by 1.61, 3.66, and 6.60 dB at 4.8 s, and by 4.57, 5.98, and 7.07 dB at 96 s. The larger long-horizon gains across all three families show that structured action conditioning also benefits extended autoregressive prediction.

Table 4: SSC ablation against zero-padded joint training. Short-horizon (4.8 s) and autoregressive long-horizon (96 s) visual prediction metrics are reported separately for each robot family, following the evaluation protocol in Appendix C.3. The SSC rows provide MiX-World's per-embodiment results. Average rows give equal weight to the three robot families; the SSC average supplies the PSNR, SSIM, and LPIPS scores in Table 1.

		Short horizon			Long horizon		
Robot family	Representation	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fixed-base bimanual	Zero-padding	22.669	0.805	0.216	14.239	0.582	0.397
	SSC	**24.281**	**0.845**	**0.154**	**18.811**	**0.720**	**0.285**
Mobile-base bimanual	Zero-padding	21.478	0.791	0.229	13.060	0.475	0.488
	SSC	**25.135**	**0.847**	**0.142**	**19.037**	**0.697**	**0.240**
Humanoid	Zero-padding	17.276	0.659	0.336	9.861	0.357	0.602
	SSC	**23.876**	**0.796**	**0.239**	**16.928**	**0.581**	**0.346**
Average	Zero-padding	20.474	0.752	0.260	12.387	0.471	0.496
	SSC	**24.431**	**0.829**	**0.178**	**18.259**	**0.666**	**0.290**

C.2 TEMPORAL SAMPLING AND CONDITIONING DETAILS

Event detection and sampling. For G1, action groups comprise the whole-body motion token and two hand commands. Manipulator groups comprise end-effector translation, rotation, grippers, base motion, and lift motion when available.

For the 50-Hz manipulator recordings, change scores are computed between consecutive raw action frames before normalization, without dividing by the time interval:

$$\begin{aligned} d_{\text{pos}}(t) &= \max_{h\in\{L,R\}} \|p_t^h - p_{t-1}^h\|_2, \\ d_{\text{rot}}(t) &= \max_{h\in\{L,R\}} \|\operatorname{wrap}_{[-\pi,\pi)}(r_t^h - r_{t-1}^h)\|_2, \\ d_{\text{grip}}(t) &= \max_{h\in\{L,R\}} |g_t^h - g_{t-1}^h|, \\ d_{\text{base}}(t) &= \|b_t - b_{t-1}\|_2, \qquad b_t = (\omega_z, v_x, v_y), \\ d_{\text{lift}}(t) &= |\ell_t - \ell_{t-1}|. \end{aligned} \tag{4}$$

Here, p, r, g, and ℓ denote end-effector position, roll–pitch–yaw angles, gripper state, and lift position, respectively. Rotation changes use component-wise wrapped RPY differences, not an SO(3) geodesic distance. Each group's threshold is its 99th percentile over valid adjacent-frame transitions in the training set. Transitions are pooled directly, without episode-level averaging or crossing valid-segment boundaries. If any available group's score is at least its threshold, both frames of the transition are marked for retention.

We construct 49-observation windows using three schedules:

Schedule	Candidate gaps (raw frames)	Start stride (raw frames)
Dense	$\{2, 3, 4\}$	72
Mixed	$\{3, 4, 5, 6, 7\}$	120
Long	$\{6, 8, 10\}$	192

One raw-frame interval is 20 ms. At each sampling step, a deterministic hash selects a candidate gap from the corresponding set. If that gap would skip a marked frame, it is shortened to end at the nearest marked frame. For mobile robots, the gap from frame i to frame j is further shortened one frame at a time while

$$\sum_{k=i}^{j-1} \frac{|\omega_z(k)|}{50} > \frac{\pi}{180},$$

subject to a minimum gap of one raw frame. The sampler also reserves enough remaining frames to complete all 49 observations within the valid segment. Actual gaps can therefore be as short as 20 ms, and the windows do not have a fixed physical duration.

Regular windows are constructed at the schedule-specific start strides above. Tail windows are included, duplicate windows are removed, and additional mixed windows cover otherwise missed event transitions. The schedules are merged without specifying fixed mixture probabilities; their final proportions depend on the constructed sampling manifest. Each window uses identical frame indices for all synchronized camera streams and their corresponding actions.

Physical-time positions. The selected raw-frame indices are $K = (k_0, \ldots, k_{48})$. We use a temporal grouping factor $q = 4$ and a reference interval $\Delta\tau_{\text{ref}} = 0.1$ s in Eq. 2, yielding 13 latent-frame anchors with $i = 0, \ldots, 12$. Uniform 10-Hz sampling gives $p_i = i$.

Intra-group timing features. For $j = 0, \ldots, 47$, the outgoing interval is $\Delta\tau_j = (k_{j+1} - k_j)/f_e$. Each valid interval is represented as

$$d_j = \left[\mu_j \frac{\Delta\tau_j}{\Delta\tau_{\text{ref}}},\ \mu_j \log\frac{\max(\Delta\tau_j, \epsilon)}{\Delta\tau_{\text{ref}}},\ \mu_j\right], \tag{5}$$

where μ_j indicates a valid outgoing interval and ϵ is a positive numerical floor for the logarithm. Padded terminal intervals have $d_j = 0$. For latent frame i, we concatenate the four interval features $d_{4i}, \ldots, d_{4i+3}$ and pass them through the physical-time conditioner. Its output projections are zero-initialized, and its outputs are added to the SSC residual δh_i^s and modulation δM_i^s.

C.3 WORLD-MODEL COMPARISON

Training corpus and representative tasks. The world-model comparison in Table 1 uses robot training data from the approximately 10,000-hour corpus described in Section 3.1. MiX-World jointly trains across the three robot families, whereas each embodiment-specific baseline uses the corpus data corresponding to its embodiment. The six tasks below are representative examples of interaction types and prediction challenges, not an exhaustive list of training tasks. They do not restrict baseline training to a separate six-task dataset:

- **Watering plants (fixed-base bimanual).** The robot positions and tilts a water container to water a plant. Predicting this interaction requires tracking the container's pose relative to the plant and the flow of water as the container tilts.
- **Opening beer bottles (fixed-base bimanual).** The robot manipulates a beer bottle to remove its cap. The key challenges are the localized contact around the bottle opening and the transition from an attached cap to a separate object, while preserving the bottle's shape and position.
- **Retrieving and folding socks (mobile-base bimanual).** The robot takes socks out of a basket and folds them. The socks deform as they are grasped, lifted, and folded, requiring predictions to capture changing fabric shape, self-occlusion, and contact with the grippers.
- **Tidying cushions (mobile-base bimanual).** The robot moves through the room to pick up and reposition cushions. This task combines base motion with manipulation of soft objects: cushions deform under grasping and change shape as they are released onto a supporting surface.
- **Tidying cushions (locomoting humanoid).** The humanoid carries and rearranges cushions while moving through the room. Prediction must account for whole-body motion together with cushion deformation, including changes in visibility as the arms and torso occlude the carried object.
- **Making a bed (locomoting humanoid).** The humanoid adjusts and smooths the bedding. Contact with the hands produces fabric displacement, folds, and wrinkles across a large surface, making it challenging to predict both local deformation and the overall arrangement of the bedding.

Baseline training and task-wise matching. We compare MiX-World with Ctrl-World (Guo et al., 2026), Motus (Bi et al., 2026), and DreamDojo (Gao et al., 2026). Each baseline is trained separately for each embodiment using the corresponding data from the corpus in Section 3.1, yielding three checkpoints per baseline method. For every task, **the training-data amount and the number of training epochs are matched** between MiX-World and the corresponding baselines. Baseline training is not restricted to the six illustrative tasks listed above. MiX-World uses a single jointly pretrained checkpoint across all three embodiments without target-specific adaptation for this comparison.

This task-wise matching controls the training-data amount and epoch count for each task. The comparison evaluates a jointly trained model against embodiment-specific models at these matched task-level budgets.

Evaluation protocol. The short- and long-horizon evaluations comprise 72 and 42 trajectories, respectively, covering 6 tasks listed above across the three robot families. Evaluation trajectories are selected randomly from evaluation set. For each trajectory, all models receive the same initial three-view observation and action sequence. We compare their predicted observations with the corresponding reference frames at 10 fps over short (4.8 s) and long (96 s) horizons, using autoregressive rollout for the latter. All models use a three-view image resolution of 640×288 pixels (width $\times$ height). Each baseline retains the view arrangement used by its pretrained model. MiX-World places the 384×288 primary view on the left and stacks two 256×144 wrist views vertically on the right.

Metrics. PSNR, SSIM, and LPIPS are computed on predicted and reference three-view composites over the corresponding prediction horizon. Scores are averaged first over frames within each trajectory, then over trajectories within each embodiment, and finally equally across the three embodiments.

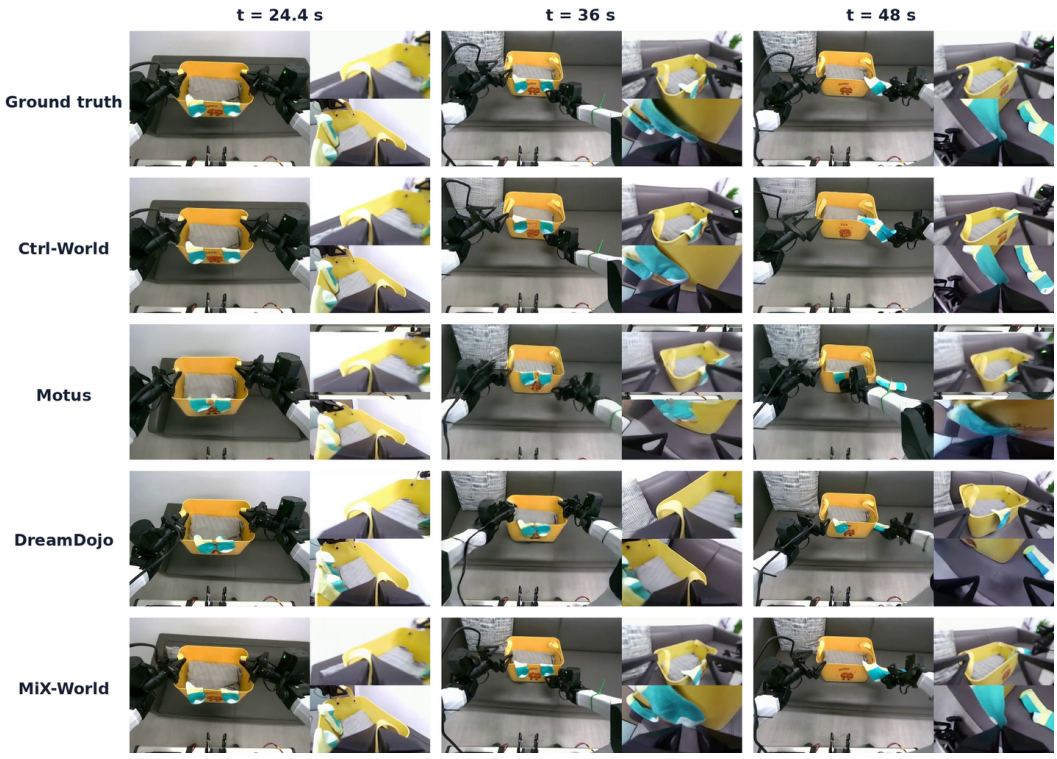


Figure 10: Ground-truth comparison of the models in Table 1 on the same mobile-bimanual sock-folding trajectory. All models receive the same initial three-view observation and action sequence. Each cell shows the primary view on the left and the two wrist views on the right at matched rollout times. The sequence covers the stage in which the robot grasps and pulls a deformable sock from the basket. MiX-World more closely matches the reference evolution of the robot, basket, and sock across the three views.

Same-architecture comparisons. We additionally compare single-embodiment and joint multi-embodiment training using the same model architecture. Table 5 compares direct predictions on fixed-base bimanual, mobile-base bimanual, and humanoid trajectories, using a separately trained single-embodiment model for each robot family. Each single-embodiment model is trained for the same number of optimization steps as MiX-World pretraining.

Table 5: Same-architecture comparisons of single-embodiment and joint multi-embodiment training. Each single-embodiment model is trained only on its corresponding robot family.

		Short horizon			Long horizon		
Robot family	Training data	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fixed-base bimanual	Single-embodiment	20.738	0.791	0.166	17.654	0.703	0.301
	Joint	24.281	0.845	0.154	18.811	0.720	0.285
Mobile-base bimanual	Single-embodiment	22.932	0.820	0.171	17.742	0.686	0.276
	Joint	25.135	0.847	0.142	19.037	0.697	0.240
Humanoid	Single-embodiment	18.208	0.682	0.291	14.308	0.557	0.438
	Joint	23.876	0.796	0.239	16.928	0.581	0.346

C.4 Qualitative ground-truth comparisons on corpus episodes

This subsection collects qualitative prediction/ground-truth comparisons on episodes from the original pretraining corpus: fixed-base bimanual sock folding and bread spreading, mobile bimanual sock folding and chip-bag retrieval, humanoid pillow transport, and flower spraying. For each panel,

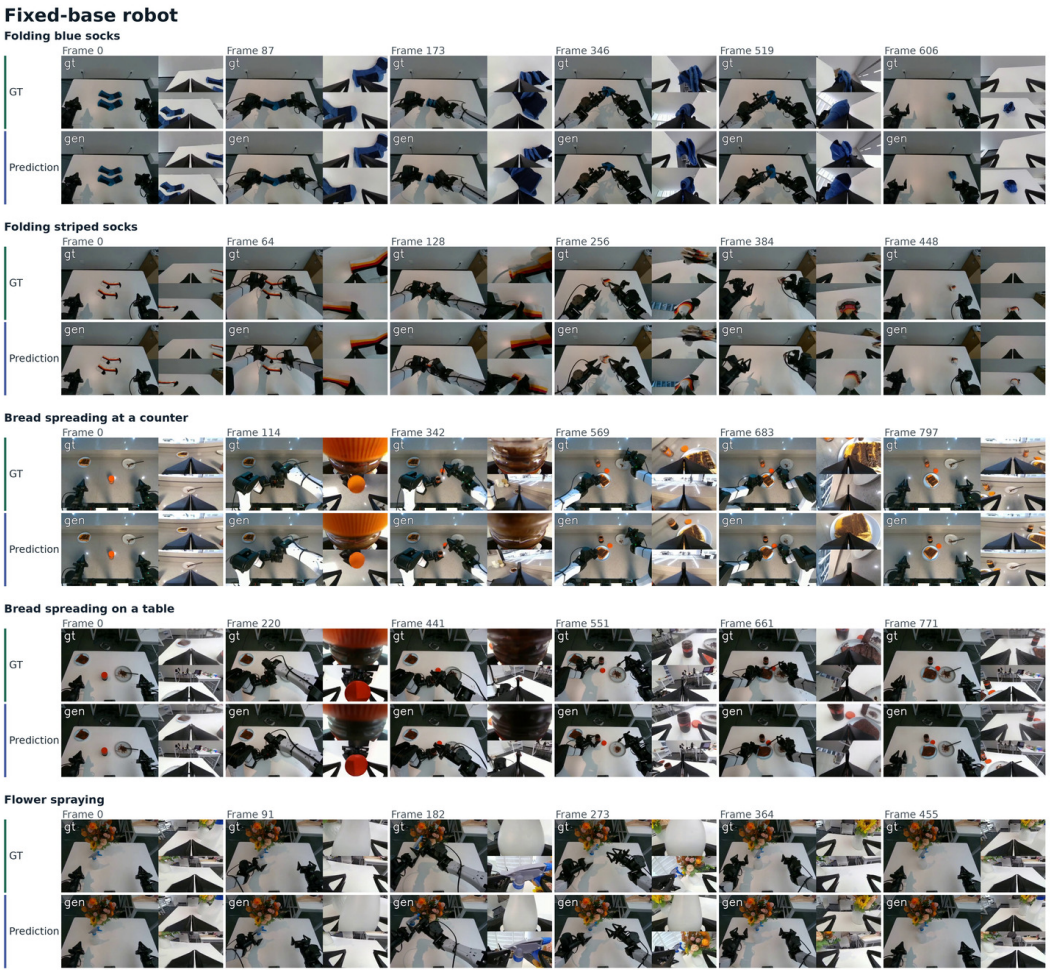


Figure 11: Qualitative comparisons against recorded ground truth on fixed-base bimanual episodes: sock folding, bread spreading, and flower spraying.

the top row shows the recorded ground truth and the bottom row shows the model prediction at identical zero-based frame indices, with camera crops fixed within each example and no retiming or interpolation. In each figure, the rows show the recorded ground truth and the model prediction across all recorded views (primary and wrist cameras). Examples cover representative manipulation phases such as approach, contact, transport, deformation, and release.

C.5 Matched comparisons and evaluation subsets

The world-model comparison in Table 1 matches training-data amounts and epoch counts for each task, following Appendix C.3. Temporal ablations use the same observation context, action inputs, and generation horizon, with 49 observations per training sample. For each paired cross-embodiment adaptation setting, the initializations use identical target examples and matched optimizer updates (Appendix E).

In addition to the full evaluation set, we report two focused subsets in Table 2. Motion-intensive clips are selected within each robot family by cumulative robot displacement, using a threshold fixed before model evaluation. Interaction-intensive clips contain visible robot-object contact or object-state change and are identified from reference trajectories before model evaluation. The temporal evaluation reported in Table 2 comprises 128 trajectories.

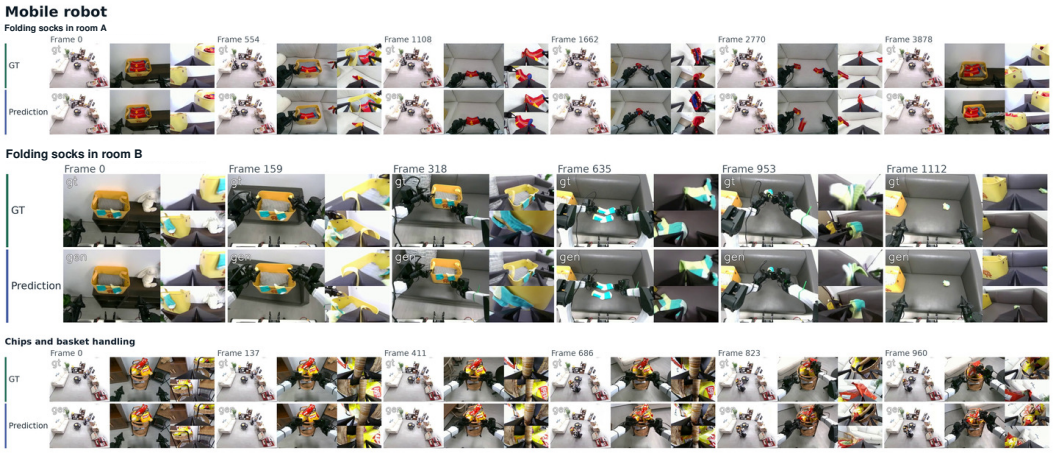


Figure 12: Qualitative comparisons on mobile-bimanual episodes: sock folding and chip-bag retrieval.

C.6 Downstream protocols

Policy evaluation and post-training use the six tasks and their corresponding robot platforms described in Appendix C.3.

We evaluate π_0 (Black et al., 2024), $\pi_{0.5}$ (Physical Intelligence, 2025), and Diffusion Policy (Chi et al., 2023) both in MiX-World and on the real robot. For each policy, the same base checkpoint is evaluated from matched held-out initial conditions. In MiX-World, the policy receives the current multi-view observations and outputs actions; the world model predicts the next observations conditioned on those actions, forming a closed loop. We generate a total of 500 trajectories with MiX-World and use them together with original data to post-train the corresponding base policies under matched policy-update budgets. All post-training success rates in Table 3 are measured on held-out real-robot trials.

C.7 Pretraining interaction taxonomy

Figure 14 expands the interaction overview of Figure 4 with the full legend: every task wedge belongs to one of 26 subtask categories, grouped into the eight interaction families of Section 3.1. Category colors follow the family hue of each wedge in the main-text overview.

D Additional Conditioning and Generalization Results

D.1 Embodiment-native conditioning controls

Figure 5 presents the fixed-base bimanual comparison in the main text. Its trajectory diagrams connect manually estimated left/right gripper centers in the generated camera images at 4.8, 9.6, 12.0, 14.4, 16.8, and 19.2 seconds. They use the same image-plane coordinates and pixel scale as the unannotated future frames, not projected 3D commands or measured end-effector poses. Nearby labels are grouped; connecting segments do not recover intervening motion. XY inversion, as an input intervention, reflects each arm's planar displacement about its initial pose, retaining the original Z commands; the model input is clipped. Right-arm hold fixes that arm's pose and gripper, and both-arm hold fixes both. The time-reversed condition reverses the temporal order of all 14 action dimensions and starts from the original rollout's generated 38.4-second end frame. The five displayed conditions use matched generation noise, and all six displayed times and one fixed tabletop crop are retained. They illustrate conditioning sensitivity rather than exact physical reversibility or trajectory-tracking accuracy.

Figure 15 provides the mobile-base counterpart, holding the initial observation and generation noise fixed while varying the input base commands. The time-reversed condition reverses the temporal order of the velocity sequence and negates its velocities; the sign-inverted condition negates velocities without reversing their order. No-op holds the remaining state fixed while zeroing all three base-velocity dimensions, and random supplies a synthetic command sequence. These are

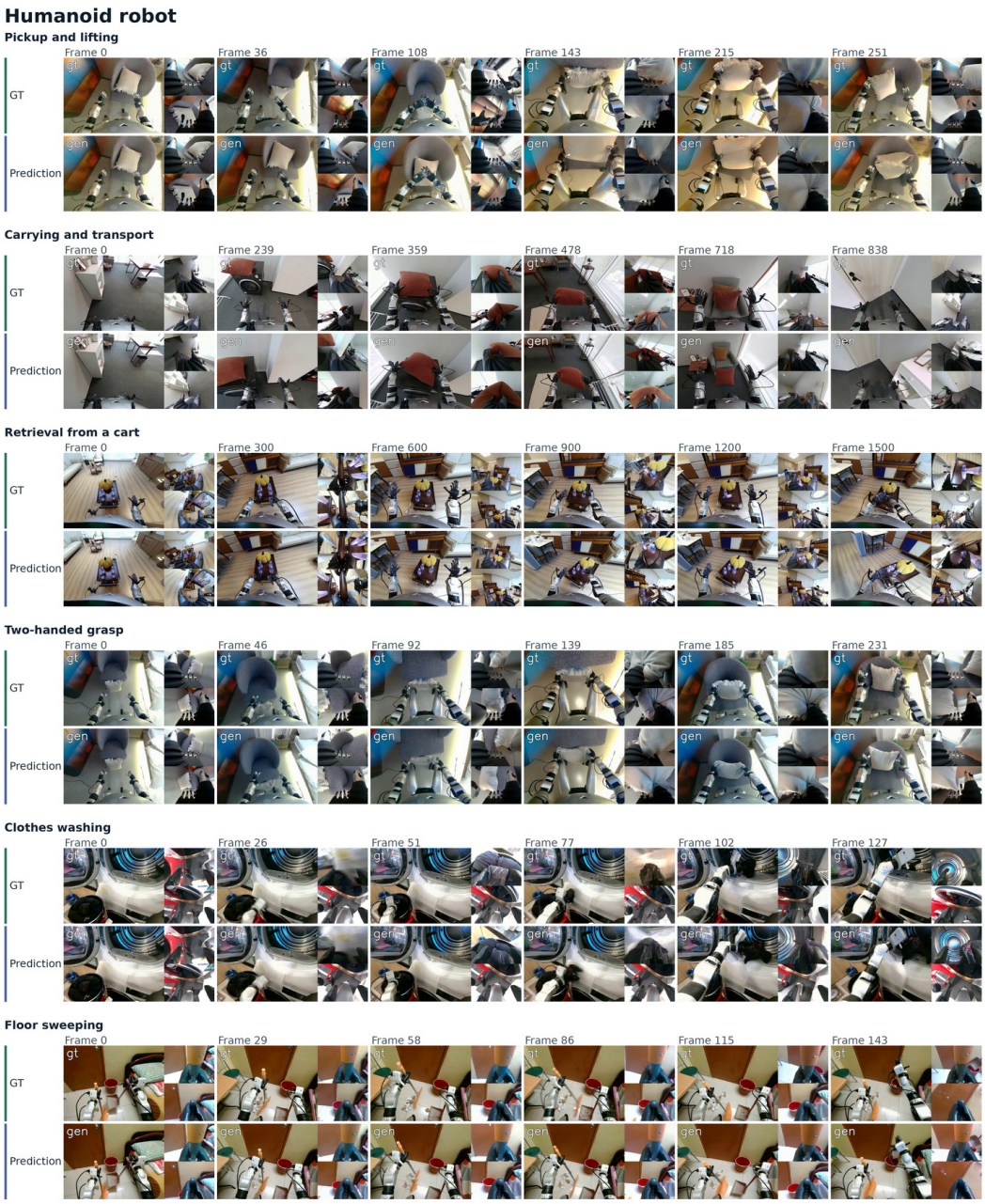


Figure 13: Qualitative comparisons on humanoid episodes.

conditioning-sensitivity examples, not measurements of exact execution accuracy. For G1, the comparison varies the recorded 64-dimensional SONIC motion-token sequence and its two hand commands across synchronized head- and wrist-camera futures.

Figure 16 extends the conditioning controls to the humanoid embodiment on the pumpkin task over a 57.6-second horizon. All six rollouts share the same initial three-camera observation, checkpoint, prompt, and generation noise (seed 42). The root quaternions and arm-joint angles are used only to construct the visualizations; the world model is conditioned on SONIC motion tokens and hand commands. In the diagrams, A/C/D/F show the source orientation sequence, B keeps the illustrated orientation fixed, and E shows the sequence in reverse. The arm schematics show the six dual-arm keyframe poses computed by forward kinematics from the 14 supplied arm-joint angles in a fixed torso frame; the reference skeleton follows the public Unitree G1 description. Generated head-camera

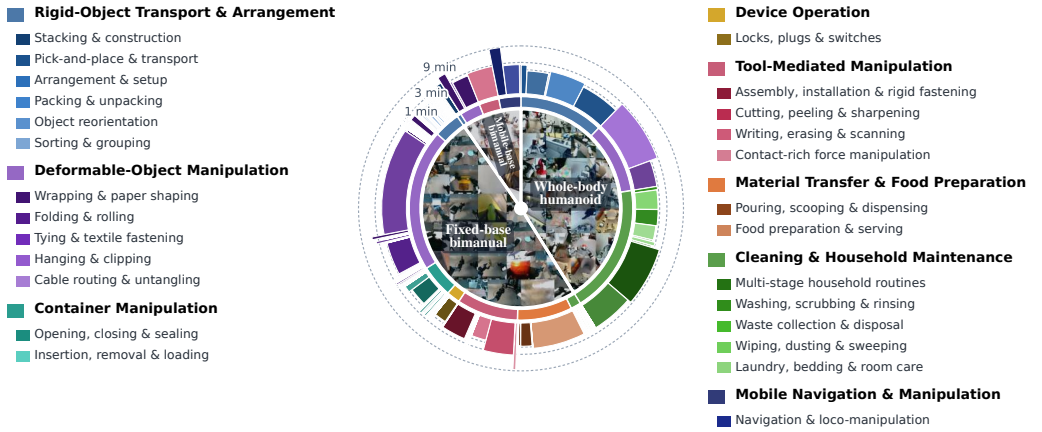


Figure 14: Subtask-level legend for the pretraining interaction overview (Figure 4). The 26 subtask categories are grouped into the eight interaction families; swatch colors match the task wedges in the overview.

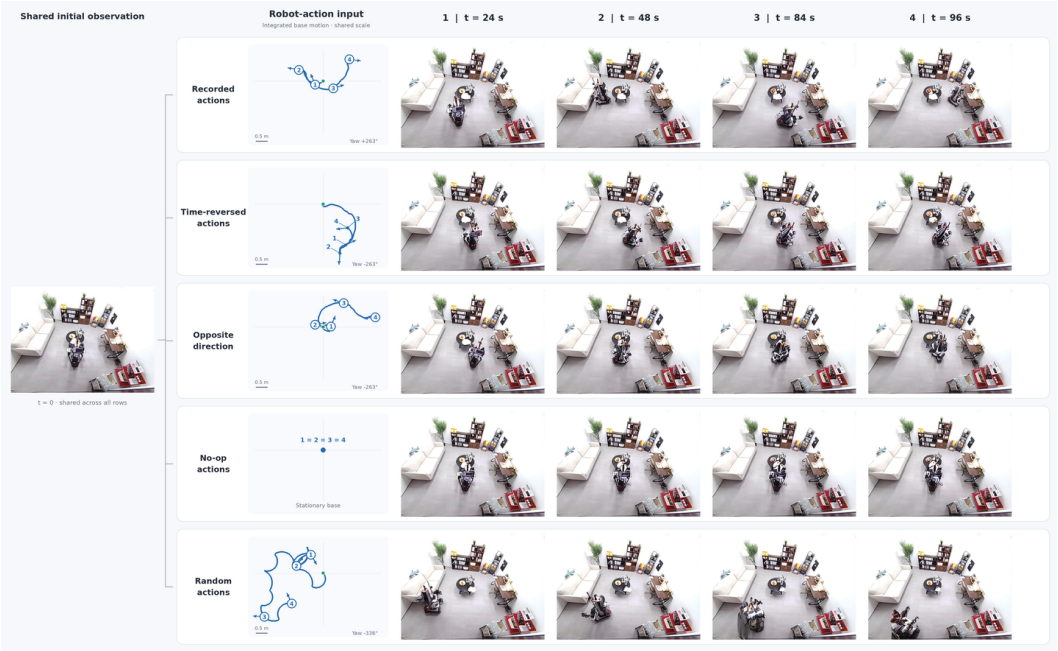


Figure 15: Action-conditioned motion prediction on the mobile bimanual embodiment. We compare five base-action sequences: original, time-reversed, sign-inverted, no-op (zero base velocity with the remaining state held fixed), and random. All rows share the same initial RGB4 observation and generation noise. Panels show the third-person ceiling view at 24, 48, 84, and 96 seconds. Trajectory diagrams integrate the input base actions on a shared scale; numbered markers identify the corresponding times. The fixed-base bimanual comparison appears in Figure 5.

frames are embedded unretouched. The room illustration is processed for privacy (the outdoor view is replaced by a flat backdrop and cart items are made semi-transparent) and is not recorded evidence.

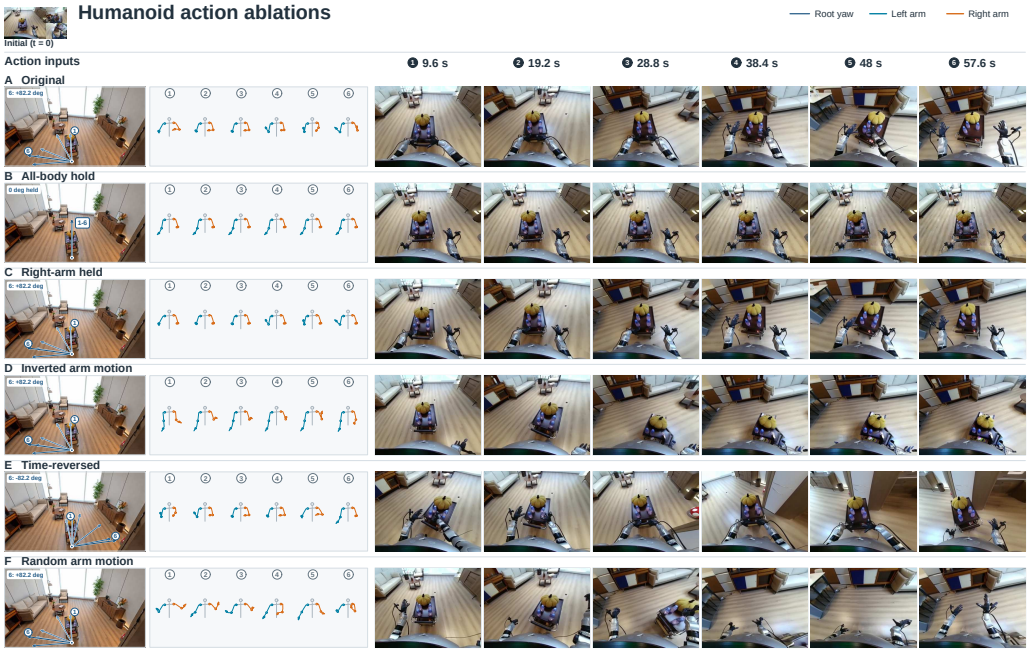


Figure 16: Humanoid action ablations on the pumpkin task from one shared initial observation. A/C/D/F share the source orientation sequence, B (all-body hold) keeps it fixed, and E (time-reversed) plays it backwards.

D.2 Constructed unseen scenes

Figure 8 in the main text shows the fixed-arm variant of this experiment. We supplement it with the mobile-bimanual variant in Figure 17 and the locomoting-humanoid variant in Figure 18: unseen scenes constructed after training through controlled synthetic edits to room appearance, layout, background, and object arrangement. These examples retain a shared source episode and platform-specific command sequence while changing one first-frame visual factor at a time. They provide a controlled qualitative test on newly constructed scene configurations. In the mobile-bimanual figure, monitoring views use a fixed $2\times$ crop; the humanoid figure shows native prediction frames without cropping.

E Cross-Embodiment Transfer

Figure 6 reports paired target-adaptation results at evaluated checkpoints. Within each family, MiX-World and the target-only model without robot pretraining use **identical target examples and matched optimizer updates**; their initializations differ by heterogeneous robot pretraining. For unseen embodiments, the same-architecture comparisons are detailed in Appendix E.1 and Table 6. The regional metrics in Figure 6(b) are defined in Appendix E.2, with ground-truth comparison examples in Figures 19 and 20.

E.1 Transfer to embodiments unseen in pretraining

The in-house adaptation families above are represented in pretraining, but their target adaptation tasks are excluded from pretraining (Section 3.3). We additionally evaluate RoboTwin 2.0 dual-arm Franka and UR5 and LIBERO (Liu et al., 2023) single-arm Franka, none of which appears in robot pretraining.

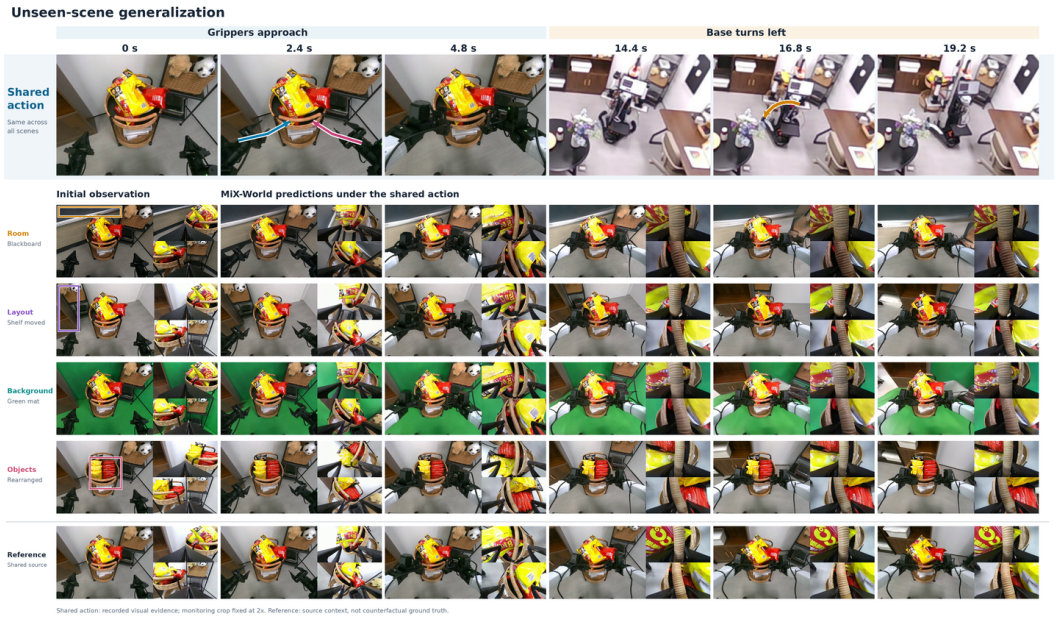


Figure 17: Constructed unseen scenes on the mobile-bimanual platform (mobile-base variant of Figure 8). Four novel configurations alter room appearance, layout, background, or object arrangement while keeping the source episode's mobile-bimanual command sequence. Each row shows the initial observation and MiX-World predictions at 2.4–19.2 seconds; the top panel evidences the shared action, and the bottom row shows the unedited source future.

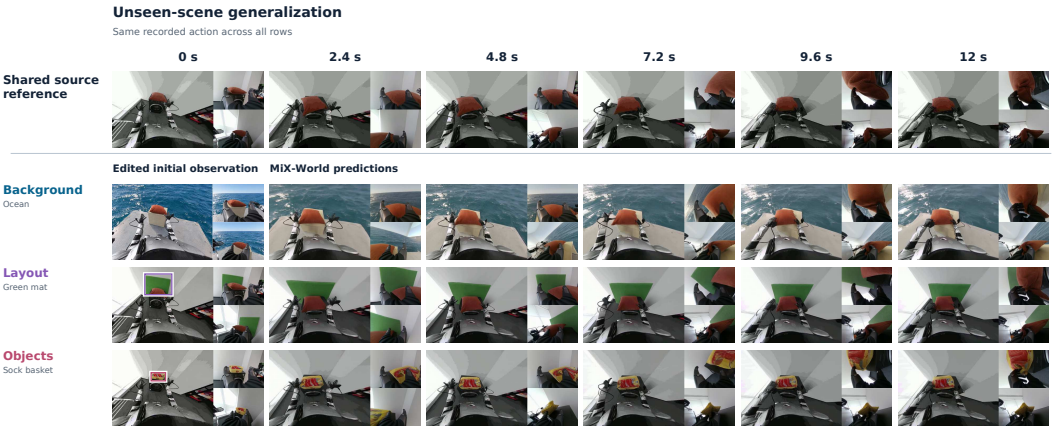


Figure 18: Constructed unseen scenes on the locomoting-humanoid platform (humanoid variant of Figure 8). Synthetic edits replace the room background with open ocean (Background), place a green mat behind the pillow (Layout), or replace the target pillow with a sock basket (Objects) while keeping the recorded pre-grasp action segment shared across rows. Each row shows the edited initial observation and MiX-World predictions at 2.4–12 seconds; the top row evidences the shared action via the unedited source future.

Controlled initialization comparison. For every paired fine-tuning setting in Table 6, the MiX-World and baseline initializations use **the same action-conditioned architecture, identical target demonstrations, and matched optimizer updates**. The same MiX-World pretrained initialization is reused across the three unseen embodiments, and paired runs are evaluated on the same test trajectories.

Tasks and target-data budgets. RoboTwin uses five tasks: lift pot, move can to pot, open laptop, stack two blocks, and hand over microphone. The measured demonstrations per task are 0, 1, 10, 50, and 100 for dual-arm Franka and 0, 1, 10, and 100 for UR5. LIBERO uses three tasks: open the middle drawer of the cabinet, put the bowl on the plate, and turn on the stove, with 0, 1, 10, and 40 demonstrations per task. Each task is evaluated on 100 held-out test trajectories, yielding 500 trajectories for each RoboTwin embodiment and 300 for LIBERO. The same test trajectories are reused across initializations and target-data budgets. One-shot therefore uses five total training trajectories for each RoboTwin embodiment and three for LIBERO.

Observations and action masks. RoboTwin composes head and two wrist views into 576×288 frames at 30 fps. LIBERO uses head and right-wrist views at 20 fps in the same frame size; only the two real camera regions are scored, excluding padding. Conditioning uses recorded end-effector poses and gripper states aligned with image timestamps. Masks use half-open, zero-based intervals in the shared 84D input, where the humanoid block (the 64-dimensional motion token and two hand commands) occupies $[0, 66)$: dual-arm Franka activates $[66, 80)$ (14 arm coordinates), UR5 activates $[66, 83)$ (both arms and three stationary-base velocities), and single-arm Franka activates $[73, 83)$ (right arm and three stationary-base velocities). The absent left arm on single-arm Franka, humanoid coordinates, and lift are masked out. Physical-zero base velocities retain frozen pretraining normalization and need not be numerically zero after normalization. For zero-shot evaluation, all normalization statistics are frozen from pretraining; fine-tuned runs fit arm statistics on their own training subset, shared between the two initializations. All three embodiments use the canonicalizer with clipping to $[-1, 1]$.

Optimization and checkpoint protocol. Zero-shot performs no target updates. One-shot uses 100 updates for all three bodies; single-arm Franka 10-demo fine-tuning also uses 100. The retained larger-budget runs use 3,000 updates on dual-arm Franka, 1,000 (10-demo) and 5,000 (100-demo) on UR5, and 5,000 (40-demo) on single-arm Franka. Short runs use a learning rate of 10^{-5} and 10 warmup updates. New runs evaluate the final checkpoint directly. For the Franka 10-demo runs, we also evaluate the 1,000-update intermediate checkpoints. The retained dual-arm Franka 100-demo results also use step 3,000, originally selected by validation. Each table row specifies its actual optimization budget.

Complete autoregressive test rollouts. Only the first real frame is provided. Each 49-frame chunk is conditioned on the preceding prediction's last frame and advances 48 frames. The final chunk holds the final action and discards padded predictions; frames 1 through $T - 1$ are scored exactly once. Sampling uses 30 denoising steps, guidance scale 1.0, and multiple paired inference seeds that vary only the initial diffusion noise; paired runs share the same seed set. Metrics report the mean and sample standard deviation across inference seeds. PSNR uses each trajectory's full-frame MSE over the real camera regions; SSIM is averaged over scored frames. Trajectories and real views receive equal weights. These metrics evaluate video prediction under recorded actions.

Results across target-data and update budgets. MiX-World improves both PSNR and SSIM in all eleven fine-tuning comparisons at matched target-data and update budgets (Table 6). With one demonstration per task and 100 updates, PSNR gains from robot pretraining are 1.74 ± 0.03 dB on dual-arm Franka, 2.85 ± 0.10 dB on dual-arm UR5, and 0.35 ± 0.13 dB on single-arm Franka. Across the measured fine-tuning budgets, seed-averaged gains range from 1.74–2.97, 2.85–3.74, and 0.35–2.09 dB, respectively; Figure 6(c) plots paired gains for selected settings.

Ten demonstrations per task suffice for MiX-World to exceed the 100-demo baseline without robot pretraining on both dual-arm embodiments. On Franka, the comparison is budget-matched: 10-demo MiX-World reaches 22.36 ± 0.03 dB and 0.7713 SSIM at 3,000 updates, versus 19.40 ± 0.13 dB and 0.7140 SSIM for the 100-demo baseline run at the same 3,000 updates. With one-tenth as many target demonstrations, MiX-World also exceeds the 100-demo baseline with one-third as many

updates on Franka (21.09 ± 0.09 versus 19.40 ± 0.13 dB) and one-fifth on UR5 (23.56 ± 0.05 versus 22.33 ± 0.22 dB). UR5 one-shot MiX-World with 100 updates also exceeds the 10-demo baseline result (20.82 versus 19.82 dB; 0.7075 versus 0.6631 SSIM). Table 7 reports all three zero-shot comparisons.

Table 6: Adaptation to unseen embodiments: MiX-World improves both metrics in all matched-budget comparisons against the baseline without robot pretraining. Both initializations use the pretrained Wan video backbone. Each row uses identical demonstrations per task and optimizer updates for both initializations. Cells report mean ± sample standard deviation over multiple inference seeds. Full-rollout metrics are evaluated on 100 held-out trajectories per task across five tasks for each dual-arm embodiment and three for single-arm Franka. Gain is the seed-averaged difference between MiX-World and the baseline, computed before rounding.

		PSNR (dB) ↑			SSIM ↑		
Demos/task	Updates	w/o robot pretrain	MiX-World	**Gain**	w/o robot pretrain	MiX-World	**Gain**
Dual-arm Franka							
1	100	16.26 ± 0.08	18.00 ± 0.11	**+1.74 ± 0.03**	0.6654 ± 0.0031	0.7059 ± 0.0014	**+0.0404 ± 0.0017**
10	1,000	18.12 ± 0.04	21.09 ± 0.09	**+2.97 ± 0.07**	0.6909 ± 0.0034	0.7518 ± 0.0020	**+0.0609 ± 0.0031**
10	3,000	19.60 ± 0.13	22.36 ± 0.03	**+2.76 ± 0.11**	0.7265 ± 0.0017	0.7713 ± 0.0004	**+0.0448 ± 0.0022**
50	3,000	19.31 ± 0.25	22.02 ± 0.10	**+2.71 ± 0.25**	0.7178 ± 0.0081	0.7726 ± 0.0014	**+0.0549 ± 0.0081**
100	3,000	19.40 ± 0.13	22.09 ± 0.03	**+2.69 ± 0.11**	0.7140 ± 0.0010	0.7621 ± 0.0016	**+0.0481 ± 0.0020**
Dual-arm UR5							
1	100	17.97 ± 0.09	20.82 ± 0.01	**+2.85 ± 0.10**	0.6275 ± 0.0020	0.7075 ± 0.0020	**+0.0800 ± 0.0023**
10	1,000	19.82 ± 0.16	23.56 ± 0.05	**+3.74 ± 0.22**	0.6631 ± 0.0045	0.7466 ± 0.0009	**+0.0836 ± 0.0054**
100	5,000	22.33 ± 0.22	25.86 ± 0.04	**+3.53 ± 0.18**	0.7281 ± 0.0065	0.7776 ± 0.0007	**+0.0495 ± 0.0058**
Single-arm Franka							
1	100	17.40 ± 0.13	17.75 ± 0.09	**+0.35 ± 0.13**	0.6728 ± 0.0044	0.6815 ± 0.0022	**+0.0086 ± 0.0030**
10	100	17.68 ± 0.03	18.21 ± 0.14	**+0.53 ± 0.12**	0.6786 ± 0.0032	0.6909 ± 0.0049	**+0.0124 ± 0.0026**
40	5,000	20.41 ± 0.08	22.49 ± 0.08	**+2.09 ± 0.04**	0.7529 ± 0.0021	0.7981 ± 0.0023	**+0.0452 ± 0.0010**

Table 7: Zero-shot prediction on unseen embodiments: no target demonstrations or updates. Both initializations use the same full-rollout test trajectories; cells report mean ± sample standard deviation over multiple inference seeds; gain is the seed-averaged difference between MiX-World and the baseline without robot pretraining.

	PSNR (dB) ↑			SSIM ↑		
Embodiment	w/o robot pretrain	MiX-World	Gain	w/o robot pretrain	MiX-World	Gain
Dual-arm Franka	13.77 ± 0.09	14.70 ± 0.06	$+0.93 \pm 0.15$	0.5443 ± 0.0016	0.6040 ± 0.0019	$+0.0597 \pm 0.0005$
Dual-arm UR5	14.22 ± 0.10	14.74 ± 0.07	$+0.52 \pm 0.14$	0.4903 ± 0.0019	0.5791 ± 0.0038	$+0.0888 \pm 0.0046$
Single-arm Franka	11.61 ± 0.41	11.72 ± 0.29	$+0.11 \pm 0.24$	0.4227 ± 0.0182	0.4407 ± 0.0092	$+0.0180 \pm 0.0117$

Table 8 reports the dual-arm Franka 100-demo full-rollout breakdown by task and camera view. Table 9 reports the corresponding breakdowns for 10/50-demo dual-arm Franka at 3,000 updates and 10/100-demo dual-arm UR5. Both tables use the same fixed checkpoints, test trajectories, and three paired inference seeds as Table 6. Each cell reports the mean and sample standard deviation across seeds; gains are computed within each paired seed before averaging. Aggregating the unrounded task/view means with equal weights recovers the corresponding means in Table 6. The total-row standard deviations are computed across seed-level totals, rather than by averaging task/view standard deviations.

Table 8: Dual-arm Franka full-rollout test results by task and view: 100 training demonstrations per task, 3,000 updates, and 100 test trajectories per task. Cells report mean ± sample standard deviation over the same three paired inference seeds as Table 6. Δ is computed as MiX-World minus the baseline without robot pretraining within each seed before summarizing across seeds.

		PSNR (dB)			SSIM		
Task	View	w/o robot pretrain	MiX-World	Δ	w/o robot pretrain	MiX-World	Δ
Lift pot	Head	19.89±0.15	21.88±0.17	+1.99±0.11	0.848±0.001	0.871±0.003	+0.023±0.003
	Left wrist	19.16±0.05	21.26±0.12	+2.10±0.16	0.710±0.001	0.753±0.002	+0.043±0.001
	Right wrist	19.91±0.19	21.80±0.40	+1.90±0.30	0.725±0.003	0.759±0.008	+0.034±0.007
Move can to pot	Head	22.30±0.21	24.42±0.27	+2.13±0.43	0.899±0.003	0.915±0.002	+0.016±0.004
	Left wrist	29.13±0.43	32.66±0.10	+3.53±0.52	0.859±0.003	0.868±0.001	+0.009±0.002
	Right wrist	25.04±0.62	27.65±0.22	+2.61±0.74	0.733±0.006	0.748±0.004	+0.015±0.010
Open laptop	Head	14.68±0.18	14.73±0.19	+0.04±0.31	0.713±0.007	0.733±0.004	+0.020±0.011
	Left wrist	20.24±0.73	26.55±0.92	+6.31±1.65	0.632±0.022	0.757±0.015	+0.125±0.033
	Right wrist	21.29±0.44	23.76±0.39	+2.48±0.43	0.658±0.012	0.680±0.008	+0.022±0.005
Stack two blocks	Head	14.33±0.08	17.25±0.13	+2.93±0.19	0.687±0.005	0.742±0.004	+0.055±0.006
	Left wrist	21.88±1.01	26.00±0.82	+4.12±0.20	0.718±0.019	0.787±0.004	+0.068±0.021
	Right wrist	19.79±0.19	23.15±0.47	+3.36±0.40	0.681±0.007	0.782±0.002	+0.101±0.005
Hand over microphone	Head	14.37±0.05	16.45±0.11	+2.08±0.08	0.657±0.012	0.728±0.002	+0.071±0.010
	Left wrist	14.46±0.16	16.32±0.12	+1.86±0.05	0.598±0.008	0.650±0.009	+0.052±0.013
	Right wrist	14.49±0.26	17.39±0.26	+2.90±0.51	0.592±0.022	0.657±0.006	+0.065±0.025
Total (views equal weight)		19.40±0.13	22.09±0.03	+2.69±0.11	0.714±0.001	0.762±0.002	+0.048±0.002

Table 9: Additional full-test-rollout results by task and view. Cells report mean ± sample standard deviation over the same three paired inference seeds as Table 6. Δ is computed as MiX-World minus the baseline without robot pretraining within each seed before summarizing across seeds.

		PSNR (dB)			SSIM		
Task	View	w/o robot pretrain	MiX-World	Δ	w/o robot pretrain	MiX-World	Δ
Dual-arm Franka, 10 demonstrations per task, 3,000 updates							
handover mic	head	14.10±0.26	16.10±0.05	+2.00±0.25	0.664±0.005	0.731±0.001	+0.067±0.005
handover mic	left wrist	15.83±0.25	16.33±0.29	+0.49±0.04	0.618±0.004	0.648±0.003	+0.030±0.008
handover mic	right wrist	15.24±0.10	16.46±0.25	+1.22±0.26	0.621±0.003	0.686±0.003	+0.065±0.004
lift pot	head	19.41±0.09	21.24±0.21	+1.84±0.27	0.836±0.000	0.865±0.001	+0.029±0.002
lift pot	left wrist	20.11±0.32	21.73±0.08	+1.61±0.35	0.737±0.010	0.770±0.001	+0.032±0.011
lift pot	right wrist	20.22±0.11	22.12±0.19	+1.90±0.28	0.741±0.003	0.771±0.004	+0.029±0.008
move can pot	head	22.51±0.13	24.76±0.16	+2.26±0.26	0.896±0.001	0.915±0.000	+0.019±0.001
move can pot	left wrist	28.20±1.14	33.25±0.09	+5.05±1.08	0.854±0.018	0.879±0.002	+0.025±0.017
move can pot	right wrist	24.34±0.32	27.50±0.15	+3.17±0.19	0.747±0.007	0.755±0.005	+0.008±0.003
open laptop	head	14.50±0.29	14.90±0.05	+0.40±0.25	0.713±0.009	0.746±0.001	+0.033±0.010
open laptop	left wrist	20.74±0.41	27.50±0.00	+6.76±0.41	0.638±0.006	0.767±0.001	+0.129±0.007
open laptop	right wrist	20.26±0.58	23.40±0.12	+3.14±0.64	0.641±0.007	0.690±0.006	+0.049±0.009
stack blocks two	head	15.11±0.27	17.60±0.06	+2.49±0.26	0.705±0.010	0.757±0.001	+0.052±0.009
stack blocks two	left wrist	22.23±0.75	27.60±0.21	+5.37±0.64	0.727±0.018	0.795±0.005	+0.068±0.013
stack blocks two	right wrist	21.16±0.29	24.92±0.11	+3.76±0.23	0.759±0.002	0.795±0.004	+0.036±0.005
Dual-arm Franka, 50 demonstrations per task							
handover mic	head	13.74±0.21	16.17±0.28	+2.43±0.08	0.631±0.015	0.732±0.006	+0.101±0.011
handover mic	left wrist	15.17±0.37	16.21±0.12	+1.04±0.38	0.593±0.016	0.677±0.002	+0.083±0.018
handover mic	right wrist	14.83±0.17	16.00±0.25	+1.17±0.10	0.604±0.019	0.674±0.003	+0.070±0.016
lift pot	head	19.67±0.13	21.86±0.26	+2.19±0.33	0.846±0.001	0.874±0.002	+0.027±0.003
lift pot	left wrist	19.66±0.19	21.08±0.23	+1.43±0.06	0.727±0.004	0.757±0.005	+0.030±0.001
lift pot	right wrist	20.33±0.17	22.54±0.14	+2.21±0.15	0.737±0.006	0.775±0.002	+0.038±0.005
move can pot	head	21.93±0.45	24.77±0.32	+2.83±0.16	0.892±0.006	0.917±0.001	+0.025±0.005
move can pot	left wrist	27.04±1.00	33.06±0.08	+6.03±1.07	0.842±0.010	0.872±0.001	+0.030±0.010
move can pot	right wrist	25.11±0.51	27.56±0.28	+2.45±0.77	0.731±0.005	0.754±0.002	+0.023±0.005
open laptop	head	14.57±0.14	14.97±0.09	+0.40±0.06	0.721±0.001	0.747±0.001	+0.025±0.002
open laptop	left wrist	20.68±1.79	24.87±0.07	+4.19±1.74	0.648±0.059	0.771±0.002	+0.123±0.060
open laptop	right wrist	18.55±0.53	22.29±0.10	+3.74±0.47	0.608±0.014	0.692±0.003	+0.084±0.013
stack blocks two	head	14.82±0.14	17.10±0.20	+2.29±0.07	0.706±0.006	0.750±0.003	+0.044±0.006
stack blocks two	left wrist	22.33±0.72	27.43±0.52	+5.10±0.52	0.731±0.006	0.806±0.009	+0.074±0.015
stack blocks two	right wrist	21.22±0.23	24.42±0.71	+3.20±0.83	0.748±0.008	0.793±0.008	+0.045±0.012
Dual-arm UR5, 10 demonstrations per task							
handover mic	head	19.97±0.23	21.48±0.06	+1.51±0.27	0.745±0.003	0.776±0.001	+0.031±0.003
handover mic	left wrist	17.37±0.25	19.36±0.04	+1.99±0.28	0.598±0.015	0.697±0.003	+0.100±0.012
handover mic	right wrist	16.69±0.26	19.14±0.14	+2.45±0.39	0.620±0.015	0.664±0.002	+0.044±0.016
lift pot	head	24.92±0.08	27.86±0.23	+2.94±0.16	0.873±0.001	0.893±0.001	+0.020±0.002
lift pot	left wrist	19.34±0.16	19.68±0.20	+0.33±0.26	0.685±0.007	0.691±0.002	+0.006±0.009
lift pot	right wrist	19.09±0.04	20.26±0.57	+1.17±0.53	0.689±0.002	0.709±0.008	+0.020±0.008
move can pot	head	25.71±0.07	28.39±0.12	+2.68±0.12	0.888±0.002	0.913±0.001	+0.026±0.002

Table 9: Additional full-test-rollout results by task and view (continued).

Task	View	PSNR (dB)			SSIM		
		w/o robot pretrain	MiX-World	Δ	w/o robot pretrain	MiX-World	Δ
move can pot	left wrist	20.31 ± 0.84	28.94 ± 0.11	$+8.63 \pm 0.89$	0.584 ± 0.029	0.798 ± 0.003	$+0.214 \pm 0.029$
move can pot	right wrist	23.81 ± 0.42	28.88 ± 0.05	$+5.06 \pm 0.43$	0.703 ± 0.018	0.761 ± 0.003	$+0.058 \pm 0.020$
open laptop	head	19.97 ± 0.27	20.82 ± 0.11	$+0.86 \pm 0.17$	0.736 ± 0.024	0.790 ± 0.003	$+0.054 \pm 0.024$
open laptop	left wrist	18.51 ± 0.93	27.89 ± 0.04	$+9.38 \pm 0.90$	0.500 ± 0.016	0.743 ± 0.003	$+0.242 \pm 0.018$
open laptop	right wrist	18.69 ± 0.53	22.86 ± 0.20	$+4.17 \pm 0.40$	0.577 ± 0.019	0.664 ± 0.005	$+0.086 \pm 0.018$
stack blocks two	head	19.32 ± 0.04	22.22 ± 0.09	$+2.90 \pm 0.10$	0.702 ± 0.001	0.771 ± 0.001	$+0.069 \pm 0.001$
stack blocks two	left wrist	18.15 ± 0.26	23.41 ± 0.34	$+5.25 \pm 0.32$	0.508 ± 0.024	0.697 ± 0.003	$+0.189 \pm 0.022$
stack blocks two	right wrist	15.44 ± 0.26	22.22 ± 0.19	$+6.78 \pm 0.45$	0.537 ± 0.016	0.630 ± 0.005	$+0.094 \pm 0.020$
Dual-arm UR5, 100 demonstrations per task							
handover mic	head	22.86 ± 0.30	25.10 ± 0.08	$+2.24 \pm 0.32$	0.798 ± 0.006	0.824 ± 0.001	$+0.026 \pm 0.007$
handover mic	left wrist	21.05 ± 0.14	23.41 ± 0.10	$+2.36 \pm 0.08$	0.731 ± 0.009	0.757 ± 0.005	$+0.027 \pm 0.005$
handover mic	right wrist	19.48 ± 0.34	22.25 ± 0.23	$+2.76 \pm 0.14$	0.668 ± 0.010	0.727 ± 0.004	$+0.060 \pm 0.009$
lift pot	head	26.57 ± 0.06	29.39 ± 0.09	$+2.82 \pm 0.12$	0.893 ± 0.001	0.909 ± 0.001	$+0.015 \pm 0.001$
lift pot	left wrist	20.96 ± 0.10	22.47 ± 0.24	$+1.51 \pm 0.33$	0.717 ± 0.004	0.749 ± 0.008	$+0.032 \pm 0.007$
lift pot	right wrist	20.72 ± 0.06	22.59 ± 0.05	$+1.87 \pm 0.03$	0.718 ± 0.002	0.755 ± 0.001	$+0.037 \pm 0.003$
move can pot	head	27.24 ± 0.19	31.13 ± 0.24	$+3.89 \pm 0.12$	0.902 ± 0.002	0.926 ± 0.001	$+0.024 \pm 0.001$
move can pot	left wrist	26.18 ± 0.31	30.02 ± 0.24	$+3.84 \pm 0.47$	0.775 ± 0.001	0.805 ± 0.004	$+0.030 \pm 0.004$
move can pot	right wrist	23.15 ± 0.67	29.90 ± 0.18	$+6.75 \pm 0.85$	0.685 ± 0.006	0.765 ± 0.006	$+0.080 \pm 0.012$
open laptop	head	21.94 ± 0.08	23.91 ± 0.11	$+1.98 \pm 0.11$	0.816 ± 0.005	0.839 ± 0.002	$+0.022 \pm 0.005$
open laptop	left wrist	24.11 ± 0.37	29.58 ± 0.18	$+5.47 \pm 0.22$	0.697 ± 0.009	0.766 ± 0.002	$+0.069 \pm 0.010$
open laptop	right wrist	17.28 ± 1.05	24.48 ± 0.14	$+7.20 \pm 1.18$	0.479 ± 0.042	0.686 ± 0.004	$+0.207 \pm 0.045$
stack blocks two	head	21.87 ± 0.25	24.44 ± 0.10	$+2.57 \pm 0.35$	0.768 ± 0.007	0.797 ± 0.001	$+0.029 \pm 0.007$
stack blocks two	left wrist	21.45 ± 0.05	25.20 ± 0.18	$+3.75 \pm 0.14$	0.672 ± 0.002	0.712 ± 0.004	$+0.040 \pm 0.004$
stack blocks two	right wrist	20.12 ± 0.84	24.08 ± 0.31	$+3.96 \pm 1.16$	0.603 ± 0.024	0.648 ± 0.002	$+0.045 \pm 0.026$

E.2 Action and interaction metrics

Figure 6(b) complements full-frame prediction metrics with errors in the robot-arm and manipulated-object regions for the paired target-adaptation comparison. Both metrics compare predictions with recorded ground truth using reference-derived masks shared between methods. Figures 19 and 20 show the ground truth, target-only prediction, and MiX-World prediction with the same masks at the displayed frames.

Arm MSE. We evaluate action fidelity in the head-camera view over all predicted frames after the conditioning context. SAM 2 (Ravi et al., 2024) is prompted on the shared initial frame to segment the visible robot arms and grippers, and its video predictor propagates the masks through the reference sequence. The visually audited reference masks are reused for both predictions. Let G_t and P_t denote reference and predicted RGB frames scaled to $[0, 1]$, and let M_t be the union of the reference arm and gripper regions. The per-frame error is

$$e_t^{\mathrm{arm}} = \frac{\sum_{x,y,c} M_t(x,y)\,(G_t(x,y,c) - P_t(x,y,c))^2}{3\sum_{x,y} M_t(x,y)}. \tag{6}$$

Arm MSE is the temporal mean of e_t^{arm}, followed by an episode mean within each target family. This GT-only mask measures error at reference arm locations; it does not independently penalize all false-positive arms outside the reference mask. On the humanoid, per-arm mask tracking is unstable in the head-camera view, so the hardware mask contains both arms with hands, forearms, and cables plus the torso shell. Figure 6(b) reports the relative error reduction over matched target-only adaptation, $100(E_{\mathrm{target}} - E_{\mathrm{MiX}})/E_{\mathrm{target}}$; raw Arm MSE remains the underlying metric. Figure 19 illustrates the computation: the same audited mask is overlaid on the ground truth and on both generated rollouts at three selected frames of each episode.

Object MSE. The key-object metric replaces the arm mask in Eq. (6) with an audited reference mask O_t of the manipulated object in the head-camera view. For a fixed set of scored frames $\mathcal{T}$ with nonempty masks,

$$E^{\mathrm{object}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \frac{\sum_{x,y,c} O_t(x,y)\,(G_t(x,y,c) - P_t(x,y,c))^2}{3\sum_{x,y} O_t(x,y)}. \tag{7}$$

The same object masks and temporal window must be used for both methods; per-episode errors are averaged within each family before computing $100(1 - E_{\mathrm{MiX}}^{\mathrm{object}}/E_{\mathrm{target}}^{\mathrm{object}})$. This is a region-specific visual fidelity measure, not a task-success or causal-consistency score. Figure 20 shows

the corresponding computation examples with the object masks overlaid on the ground truth and both generated rollouts. SAM2.1 Hiera Small masks are propagated on GT videos and reused across methods and checkpoints; RGB values are normalized to $[0, 1]$. For fixed-base robots, the object is the manipulated white bag, scored from first contact through frame 384 (chunk 8); for mobile-base robots, it is the list sheet, scored from first contact through the actual video end. Frame 0 is excluded. For each of the fixed-base, mobile-base, and G1 families, results are averaged equally over 40 evaluation trajectories. On the humanoid, the object is the throw pillow, scored over its carry, placement, and release window. Mask review is agent- and vision-model-assisted, not human-annotator review.

Values and aggregation for panel (b). Table 10 lists the per-family values behind Figure 6(b), which uses checkpoints at 0.25 training epochs for both initializations. Within each family, relative improvement is computed as $100(S_{\mathrm{MiX}}/S_{\mathrm{base}} - 1)$ for SSIM and $100(1 - E_{\mathrm{MiX}}/E_{\mathrm{base}})$ for MSE; the three families are then averaged equally per metric, yielding 55.08%, 77.00%, and 66.89%, displayed as 55%, 77%, and 67% in the figure. For each of the fixed-base, mobile-base, and G1 families, uncertainty in Object MSE is estimated using 10,000 paired trajectory-bootstrap resamples. Each resample draws 40 trajectories with replacement, using the same trajectory indices for the baseline and MiX-World, and recomputes the mean errors and relative reduction. The object evaluations use family-specific temporal windows, so their average is descriptive rather than an estimate under a common evaluation horizon.

Table 10: Regional-metric values underlying Figure 6(b). Each entry compares the baseline without robot pretraining and MiX-World at 0.25 training epochs on the unseen-task evaluation of each pretraining family. Improvement within a family is $100(S_{\mathrm{MiX}}/S_{\mathrm{base}} - 1)$ for SSIM and $100(1 - E_{\mathrm{MiX}}/E_{\mathrm{base}})$ for MSE. The mean row averages the three families equally per metric; these are the 55%, 77%, and 67% displayed in the figure.

	SSIM ↑			Arm MSE ↓			Object MSE ↓		
Family	w/o robot pretrain	MiX-World	Impr. (%)	w/o robot pretrain	MiX-World	Impr. (%)	w/o robot pretrain	MiX-World	Impr. (%)
Fixed-base	0.4561	0.6921	51.76	0.1412	0.0180	87.23	0.0468	0.0136	70.98
Mobile-base	0.4137	0.6640	60.49	0.0707	0.0191	72.95	0.1211	0.0513	57.62
G1	0.4194	0.6417	53.00	0.0685	0.0200	70.81	0.0808	0.0226	72.08
Mean	–	–	**55.08**	–	–	**77.00**	–	–	**66.89**

For all three families, Object MSE uncertainty is estimated with 10,000 paired trajectory-bootstrap resamples. Object evaluations use family-specific temporal windows, so the Object MSE mean is descriptive rather than an estimate under a common evaluation horizon. The G1 Arm MSE mask covers both arms and the torso shell, as in Eq. (6).

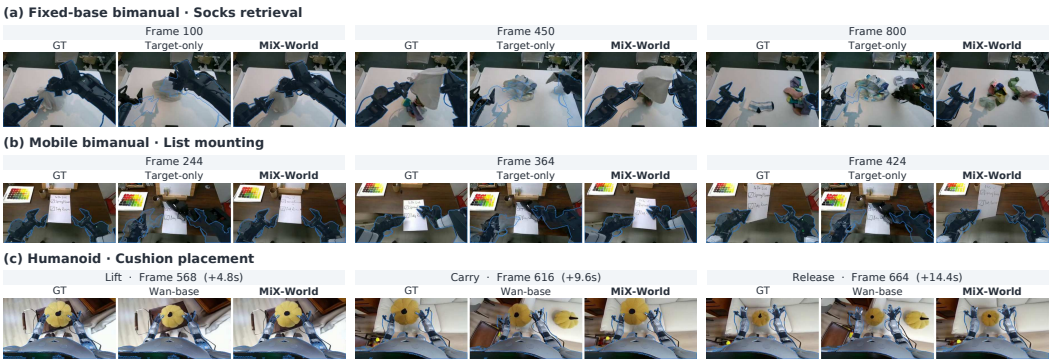


Figure 19: Selected frames from the evaluation trajectories used to compute Arm MSE. Rows show (a) fixed-base socks retrieval (bag interaction), (b) mobile list mounting, and (c) humanoid cushion placement. Each triplet compares ground truth, target-only adaptation without robot pretraining, and MiX-World-initialized target adaptation at the labeled source-frame index. The same ground-truth mask used for metric computation is overlaid in blue on all three images, with a shared crop. The G1 mask includes both arms and the torso.

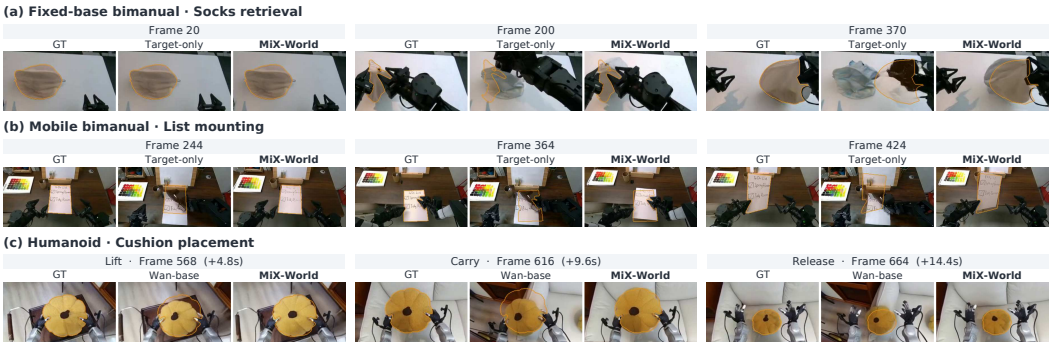


Figure 20: Selected frames from the evaluation trajectories used to compute Object MSE on the manipulated object of each task (bag, to-do-list sheet, and throw pillow). Rows and triplet order follow Figure 19; the same ground-truth object mask used for metric computation (orange) is overlaid on all three panes of a triplet as a translucent fill with a saturated outline. Frames are cropped around the mask region. The humanoid row shares the episode and checkpoint pair of Figure 19; GT frames 568, 616, and 664 show lifting, carrying, and release, respectively.

F Temporal Modeling Ablations

F.1 Original frames for the temporal comparison

Figure 21 shows the original four-view frames underlying Figure 7 at 0, 7.34, and 15.66 seconds, comparing ground truth, fixed-rate sampling, and EAS with TA-RoPE. The views are separated from their original composite without retouching or background replacement.



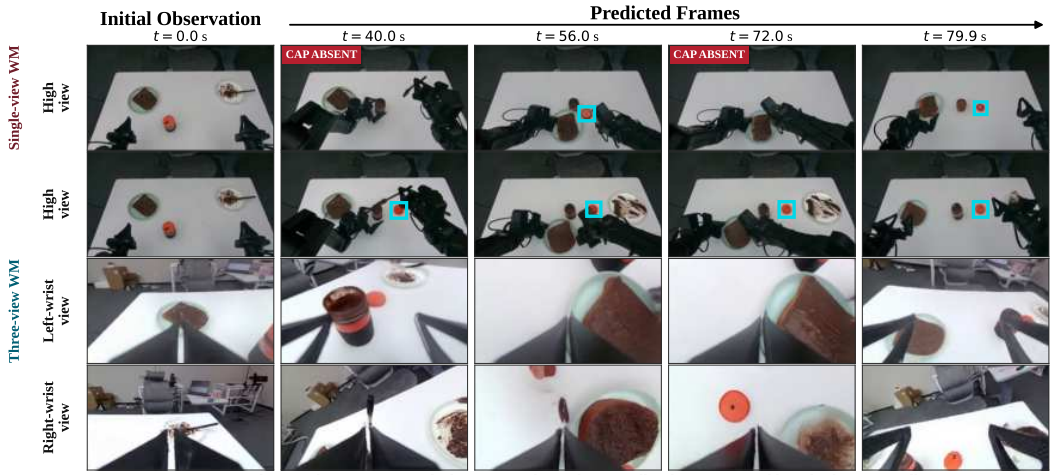
Figure 21: Original four-view frames underlying Figure 7. Rows compare ground truth, fixed-rate sampling, and Ours; columns show elapsed times of 0, 7.34, and 15.66 seconds. Each cell collects the four views of one method at one physical time—ceiling (top left), head (top right), and left and right wrist (bottom)—allowing comparison at the same physical time across all four cameras. Both models receive the same initial observation, and the ceiling view shows the difference in motion progression with the robot-mounted views alongside.

G AUXILIARY MULTI-VIEW RESULTS

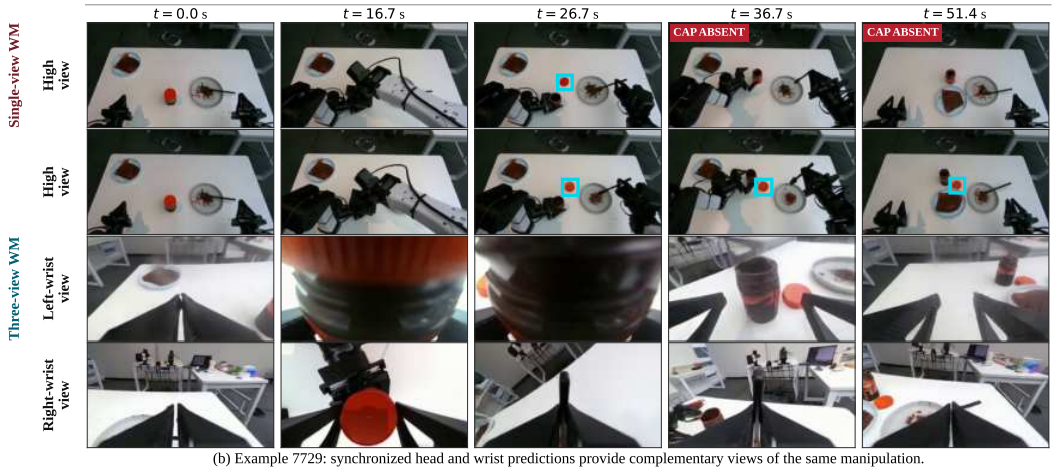
MiX-World jointly predicts synchronized camera views using the shared model trained on both three-view and four-view trajectories. The three-view configuration used in the main visual prediction benchmark consists of a primary camera and two wrist cameras. G1 uses a head-mounted primary camera, while a top-mounted third-person primary camera is used in a subset of mobile-base YAM recordings. The temporal comparison in Figures 7 and 21 uses a four-view example comprising ceiling, head, and left/right wrist cameras to facilitate visualization of base motion. Both compared models jointly predict all four views; the ceiling view provides a clear view of the robot's position and orientation over time.

We use two long-horizon bread-spreading rollouts to diagnose how joint multi-view prediction affects key-object persistence (Fig. 22). For example 7432, Presence Recall on `cam_high` is 35.18% for the single-view model and 96.97% for the three-view model ($+61.79$ percentage points). Presence Recall measures the fraction of evaluable reference-visible frames in which the generated cap is detected.

This object-centric diagnostic complements full-frame metrics. Over the second half of example 7432, the single-view rollout obtains better PSNR, SSIM, and LPIPS ($15.73/0.735/0.312$ versus $14.36/0.700/0.337$), yet repeatedly drops the cap. The comparison exposes persistent object state beyond global pixel similarity. Both the single-view and three-view models are evaluated at step 2400. Both examples (7432 and 7729) are from the test set.



(a) Example 7432: the cap intermittently disappears in the single-view rollout but remains visible in the three-view rollout.



(b) Example 7729: synchronized head and wrist predictions provide complementary views of the same manipulation.

Figure 22: Single-view versus synchronized three-view prediction on two bread-spreading rollouts. Columns are aligned rollout times; rows show the single-view high camera and the jointly generated high, left-wrist, and right-wrist cameras. Cyan boxes mark a detected cap and red labels mark disappearance in non-occluded high-camera frames. The occluded t=16.7 s frames in (b) are unmarked. In both examples, the three-view rollout preserves the cap after it disappears from the single-view rollout.