Hi-WM: Human-in-the-World-Model for Scalable Robot Post-Training

Yaxuan Li[**], Zhongyi Zhou[**], Yefei Chen[**], Yanjiang Guo[†],
Jiaming Liu[‡], Shanghang Zhang[‡], Jianyu Chen[†], and Yichen Zhu[§]
[*]Current Robotics [†]Tsinghua University [‡]Peking University [§]University of Toronto
[*]Equal contribution.

***Abstract*—Post-training is essential for turning pretrained generalist robot policies into reliable task-specific controllers, but existing human-in-the-loop pipelines remain tied to physical execution: each correction requires robot time, scene setup, resets, and operator supervision in the real world. Meanwhile, action-conditioned world models have been studied mainly for imagination, synthetic data generation, and policy evaluation. We propose Human-in-the-World-Model (Hi-WM), a post-training framework that uses a learned world model as a reusable corrective substrate for failure-targeted policy improvement. A policy is first rolled out in closed loop inside the world model; when the rollout becomes incorrect or failure-prone, a human intervenes directly in the model to provide short corrective actions. Hi-WM caches intermediate states and supports rollback and branching, allowing a single failure state to be reused for multiple corrective continuations and yielding dense supervision around behaviors that the base policy handles poorly. The resulting corrective trajectories are then added back to the training set for post-training. We evaluate Hi-WM on three real-world manipulation tasks spanning both rigid and deformable object interaction, and on two policy backbones. Hi-WM improves real-world success by 37.9 points on average over the base policy and by 19.0 points over a world-model closed-loop baseline, while world-model evaluation correlates strongly with real-world performance (r = 0.953). These results suggest that world models can serve not only as generators or evaluators, but also as effective corrective substrates for scalable robot post-training.**

I. Introduction

Large-scale robot pretraining has produced generalist policies that transfer across tasks, objects, and language instructions. Yet broad pretraining alone is rarely sufficient for deployment. In concrete scenes, policies still fail under contact-rich dynamics, environment-specific variation, and long-horizon error accumulation. In practice, post-training is what turns a generalist policy into a reliably deployable, task- or scene-specialized policy.

The most effective post-training pipelines are still on-policy and corrective: let the current policy act, observe where it breaks, collect targeted supervision on those failure states, and update the policy accordingly. However, this loop remains tied to the physical world. Every rollout requires real robots, real scenes, human monitoring, and often manual reset. As coverage expands to more tasks, scenes, checkpoints, and long-tail failure modes, the marginal cost of post-training grows rapidly. For generalist policies, the bottleneck is therefore not only data efficiency, but deployment economics.

World models suggest a different substrate for post-training. Prior work has used learned dynamics primarily for imagination, synthetic data generation, policy optimization, or post-hoc evaluation. In contrast, we study the world model as an interactive corrective workspace for post-training. Our key observation is that the most valuable supervision often lies near the learner's own failure states: not in generic successful rollouts, but in short corrective continuations that recover from imminent failure. If such corrections are collected only in the physical world, each intervention remains tied to hardware time, scene reset, and sequential deployment.

We therefore propose **Human-in-the-World-Model (Hi-WM)**, a post-training framework in which a pretrained policy is rolled out inside an action-conditioned world model, and a human intervenes only when the rollout becomes incorrect, uncertain, or failure-prone. Instead of treating the world model as a generator or evaluator alone, Hi-WM uses it as a reusable corrective substrate: failure states can be cached, rewound, and branched into multiple alternative continuations, yielding dense recovery supervision without repeated real-world execution. The resulting corrective trajectories are then used for post-training, e.g., with imitation learning or reinforcement learning. In this way, Hi-WM preserves the benefits of targeted human correction while changing the cost structure of post-training from a hardware-bound loop to a reusable interactive process.

We instantiate Hi-WM on three real-world manipulation tasks spanning rigid and deformable object interaction. Across these tasks, we study whether the world model is faithful enough to support intervention, whether world-model-collected corrective trajectories improve real-world policy performance, and whether the benefits scale with additional intervention data that are collected from the world simulator. Our results show that Hi-WM consistently improves real-world success over both the base policy and a world-model closed-loop baseline, suggesting that the main benefit comes not simply from more simulated rollouts, but from failure-targeted human correction inside the learned environment.

Our contributions are three-fold:

- **Human corrective post-training in the world model.** We formulate post-training as targeted human correction inside an action-conditioned world model, rather than collecting corrective supervision only during physical robot execution.
- **Rollback-and-branching corrective data collection.** We

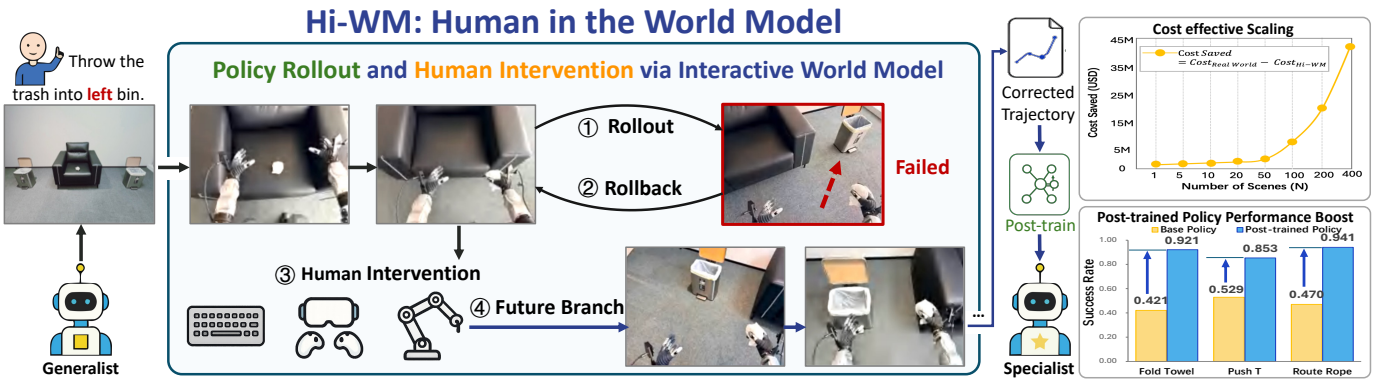

Fig. 1: Hi-WM overview. A policy rolls out in a learned world model. When failure-prone, a human provides short corrective actions. Cached states support rollback and branching, producing continuations for post-training.

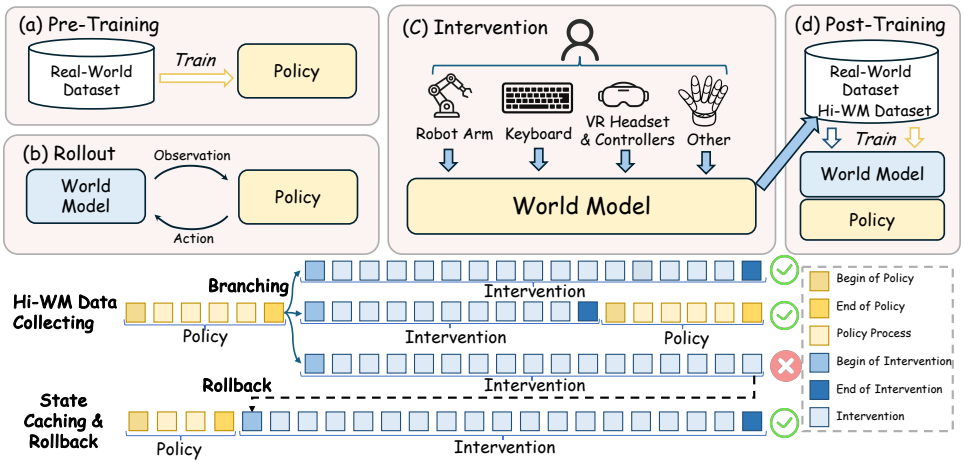

Fig. 2: **Hi-WM pipeline and rollback.** A policy rolls out inside the world model; a human intervenes when it becomes failure-prone; cached states enable rollback and branching, producing corrective continuations for post-training.

introduce a data collection mechanism that caches pre-failure states and reuses them to collect multiple corrective continuations, increasing the density and diversity of recovery supervision around failure-prone states.

- **Real-robot evidence for effectiveness and scaling.** We show on three real-world manipulation tasks and two policy backbones that corrective trajectories collected in the world model improve downstream real-world performance, while offering increasingly favorable economics as virtual intervention data scale.

II. Method

A. The Human-in-the-World-Model Paradigm

We formalize Human-in-the-World-Model (Hi-WM) as a novel post-training paradigm that shifts the human intervention loop from physical execution to an interactive world model. Unlike standard physical deployment, which is constrained by sequential and hardware-bound data collection, the Hi-WM paradigm has three practical properties that make this post-training loop scalable:

- **State caching and trajectory branching.** Because the rollout happens inside the world model, intermediate states can be cached throughout the interaction. This allows for a direct rollback to any previously failed timestep, from which the system can execute new actions to form alternative recovery branches.
- **Reset-free data collection.** Achieving reset-free operation is a crucial prerequisite for fully automated evaluation, yet it remains a formidable challenge in physical settings. By virtualizing the interaction environment, this approach removes the need for manual scene resets, naturally enabling continuous corrective interaction.
- **Hardware-agnostic remote collection.** Since human operators interact exclusively with the world model, the data collection process is completely separated from the physical robot. This allows researchers to gather corrective data remotely and in parallel, without being restricted by hardware access or location.

B. Interactive World Model

The world model is an action-conditioned visual dynamics model with a visual encoder, latent dynamics module, and visual decoder. It is conditioned directly on the 14-dimensional continuous action space of our dual-arm platform, where

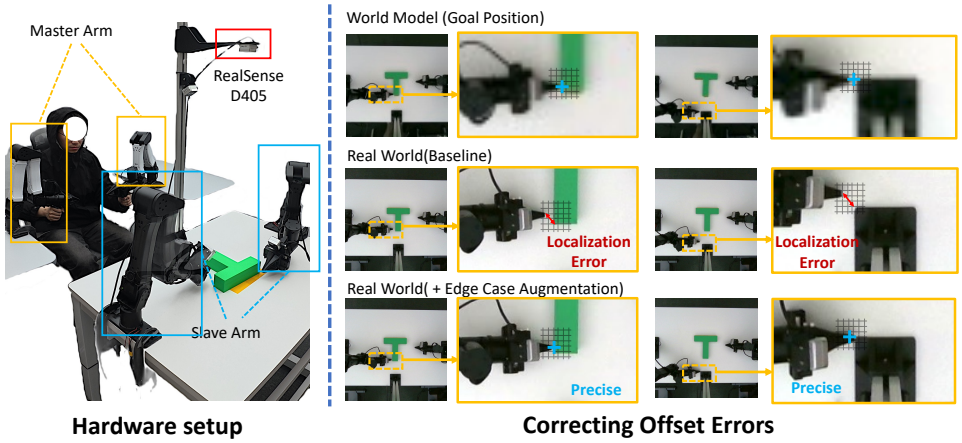

Fig. 3: **Hardware setup and offset correction. Left:** Master-slave teleoperation system with an overhead RealSense D405 camera. **Right:** Edge-case augmentation improves real-world precision by reducing localization offsets.

each arm contributes a 6-DoF end-effector pose and a 1-DoF gripper state. This avoids collapsing robot control into low-dimensional or discrete commands and preserves the precision needed for manipulation.

To make the model useful for intervention, we train it not only on successful demonstrations but also on failed and off-task trajectories. These data reduce over-optimistic predictions in which the model would otherwise hallucinate task success from roughly correct motion. We further include edge-case trajectories near workspace boundaries and precision-grasp configurations, improving world-to-real alignment in the regions where interventions are likely to occur. Additional alignment diagnostics are reported in Figure 3.

The model is used as an interactive simulator rather than as a one-shot video predictor. Its output is repeatedly fed back into the policy, and the next policy action is conditioned on the generated observation. This closed-loop use makes action faithfulness important: small deviations in predicted gripper pose, object contact, or object geometry can change when an operator intervenes. For this reason, we emphasize coverage of failure and edge states rather than optimizing only average visual fidelity on successful demonstrations.

C. Intervention Interface and Post-Training

Hi-WM decouples the human input device from the target robot. Keyboard commands, calibrated master-arm motions, and VR controller poses can all be converted by a unified action mapper into the same continuous action representation consumed by the policy and the world model.

During a rollout, Hi-WM records the policy segment before intervention, the human corrective segment, and the post-intervention continuation. When an error occurs, the operator can rewind to an earlier cached state and try an alternative correction, producing several branches from one failure. We merge these corrective segments with the original real-world demonstrations and fine-tune the policy with the same supervised post-training objective used for the base policy. Because the new data are concentrated near learner-induced failures, the update targets behaviors that generic demonstrations or self-generated successful rollouts tend to under-sample.

III. Experiments

In this section, we conduct extensive real-robot experiments to demonstrate the effectiveness and scalability of our proposed method. Specifically, we aim to answer the following core research questions. More details are shown in appendix in Section V.

- **RQ1 (Real-World alignment of Hi-WM):** Is Hi-WM sufficiently aligned with real-world execution to support reliable human intervention and data collection?
- **RQ2 (Effectiveness):** Can Hi-WM serve as an effective, policy-agnostic post-training framework?
- **RQ3 (Scalability):** How does Hi-WM scale in cost advantage and real-world policy performance as scene coverage and virtual intervention data increase?
- **RQ4 (Real-World generalization):** Does Hi-WM intervention on novel scenarios enhance the policy's real-world generalization?

A. Real-World Alignment of Hi-WM

For human intervention to be effective, the interactive world model must remain aligned with physical execution. We evaluate this alignment from three perspectives: whether model-specified target positions can be reproduced by the real robot, whether the model supports stable interpolation within the data-covered workspace, and whether world-model task performance matches real-world performance.

1) Repeated positioning accuracy: We run lower-level diagnostics to test whether model-specified target states can be reproduced by the real robot. Boundary regions and motion-limit regions are the most error-prone, so we augment training with edge-case data near these states. The added data reduce positioning offsets and improve visual fidelity metrics, as reported in Figure 3.

2) Interpolation within the covered workspace: Accurate reproduction of a small set of observed points is not enough. A useful interactive world model should also support stable interpolation within the covered workspace. We further test a

TABLE I: **Impact of edge-case data scale on visual fidelity.** We measure image prediction quality with PSNR (↑), SSIM (↑), and LPIPS (↓). Percentages indicate the fraction of edge-case trajectories added for augmentation relative to the complete edge-case dataset.

Ratio (%)	PSNR ↑	SSIM ↑	LPIPS ↓
0 (Base)	18.50	0.815	0.152
20	19.82	0.853	0.124
50	21.15	0.901	0.098
100 (Full)	**22.53**	**0.942**	**0.055**

TABLE II: **Real-world success rates (%) under the standard task setting.** Our Hi-WM achieves the best result under matched post-training budgets.

Method	Policy	Success Rate (%)		
		Fold Towel	Push T	Route Rope
Base	DP	42.1	52.9	47.0
	Pi0	55.3	76.5	64.7
WM-CL	DP	76.3	64.7	70.6
	Pi0	78.9	79.4	82.4
Ours	DP	**92.1**	**85.3**	**94.1**
	Pi0	**97.4**	**97.1**	**100.0**

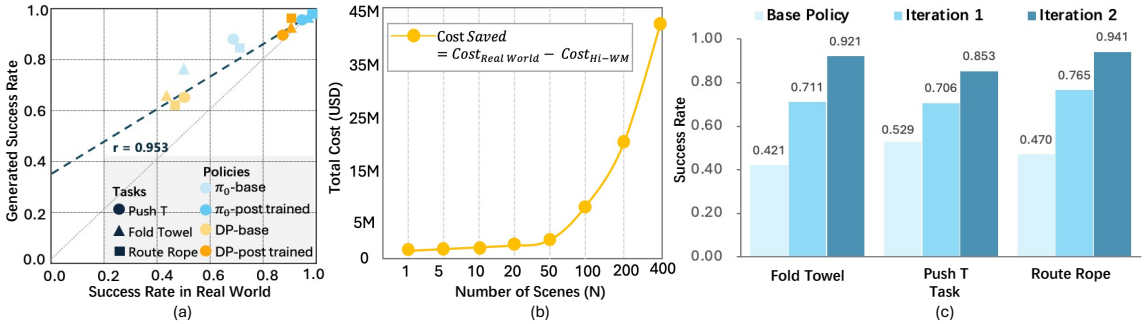

Fig. 4: **Correlation, cost savings, and intervention-data scaling. (a):** World-model success strongly correlates with real-world success. **(b):** Hi-WM yields increasing cumulative cost savings as scene coverage grows. **(c):** Real-world success improves consistently with additional virtual intervention data.

continuous ten-minute Push-T interaction and observe stable robot-object motion through interior states, suggesting that the model interpolates within the covered workspace rather than only replaying sparse anchors.

3) Task level evaluation consistency: We test whether task performance in the world model correlates with real-world performance. As shown in Figure 4(a), the two success rates are strongly correlated, with a Pearson correlation coefficient of $r = 0.953$. This indicates that the world model is predictive at the task level and can serve as a reliable proxy for policy evaluation, helping operators identify when intervention is needed and supporting high-quality corrective data collection for post-training.

B. Policy Performance via Hi-WM

We evaluate whether the resulting corrective trajectories improve downstream policy performance. We compare it against the following baselines:

Base Policy: A policy trained solely on the offline dataset without any human intervention.

World Model Closed-Loop(WM-CL): The policy is rolled out within the world model and only the generated successful trajectories are collected. The policy is then fine-tuned on these self-generated data, making this a human-free baseline.

Table 2 reports real-world success rates on the three tasks under matched post-training budgets for two policies, DP [11] and π_0 [8]. Hi-WM achieves the best result in all settings, showing that its benefit is not tied to a particular policy. Averaged across settings, our method improves success by 37.9 percentage points over the base policy and by 19.0 points over the world-model closed-loop baseline. This baseline isolates the effect of adding successful rollouts generated by the current policy without human correction; although it improves over the base policy in all six settings, it remains consistently below Hi-WM.

We attribute this gap to the type of supervision collected by each method. The closed-loop baseline mainly reinforces behaviors already available to the current policy. In contrast, Hi-WM introduces human correction near emerging failures, concentrating supervision around failure-prone states that the base policy does not handle well. Its rollback-and-branching mechanism further allows multiple corrective continuations to be collected from the same pre-failure state, providing richer recovery supervision and leading to higher real-world success after post-training.

IV. Conclusion

We presented Hi-WM, a post-training framework that uses an interactive world model as a reusable corrective substrate for robot policies. By shifting human intervention from physical execution to cached, rollbackable, and branchable model rollouts, Hi-WM collects dense recovery supervision around failure-prone states while reducing repeated robot deployment. Across three real-world manipulation tasks and two policy backbones, Hi-WM improves success over both base policies and a human-free world-model closed-loop baseline, suggesting a scalable and cost-effective route for robot post-training.

References

[1] 1X Technologies. 1X World Model — 1x.tech. https://www.1x.tech/discover/1x-world-model, 2025. [Accessed 16-05-2025].

[2] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[3] Maren Awiszus, Frederik Schubert, and Bodo Rosenhahn. World-gan: a generative model for minecraft worlds. In *2021 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2021.

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv preprint arXiv:2412.03572*, 2024.

[5] Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. In *International Conference on Learning Representations*, volume 2025, pages 56729–56763, 2025.

[6] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

[7] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

[8] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[9] 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.

[10] Xiaoyu Chen, Junliang Guo, Tianyu He, Chuheng Zhang, Pushi Zhang, Derek Cathera Yang, Li Zhao, and Jiang Bian. Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024.

[11] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[12] Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

[13] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[14] Jensen Gao, Suneel Belkhale, Sudeep Dasari, Ashwin Balakrishna, Dhruv Shah, and Dorsa Sadigh. A taxonomy for evaluating generalist robot manipulation policies. *arXiv preprint arXiv:2503.01238*, 2025.

[15] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[16] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

[17] Yanjiang Guo, Tony Lee, Lucy Xiaoyang Shi, Jianyu Chen, Percy Liang, and Chelsea Finn. Vlaw: Iterative co-improvement of vision-language-action policy and world model. *arXiv preprint arXiv:2602.12063*, 2026.

[18] D HO, J MONAS, JT REN, and C YU. 1x world model: evaluating bits, not atoms, 2025.

[19] Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning. In *Proceedings of the 5th Conference on Robot Learning*, pages 598–608, 2022.

[20] Ryan Hoque, Lawrence Yunliang Chen, Satvik Sharma, Karthik Dharmarajan, Brijen Thananjeyan, Pieter Abbeel, and Ken Goldberg. Fleet-DAgger: Interactive robot fleet learning with scalable human supervision. In *Proceedings of The 6th Conference on Robot Learning*, pages 368–380, 2023.

[21] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[22] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

[23] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng,

Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through video world models. In *Proceedings of The 9th Conference on Robot Learning*, pages 5170–5194, 2025.

[24] Snehal Jauhri, Carlos Celemin, and Jens Kober. Interactive imitation learning in state-space. In *Proceedings of the 2020 Conference on Robot Learning*, pages 682–692, 2021.

[25] Yueru Jia, Jiaming Liu, Shengbang Liu, Rui Zhou, Wanhe Yu, Yuyang Yan, Xiaowei Chi, Yandong Guo, Boxin Shi, and Shanghang Zhang. Video2act: A dual-system video diffusion policy with robotic spatio-motional modeling. *arXiv preprint arXiv:2512.03044*, 2025.

[26] Yunfan Jiang, Chen Wang, Ruohan Zhang, Jiajun Wu, and Fei-Fei Li. TRANSIC: Sim-to-real policy transfer by learning from online correction. In *Proceedings of The 8th Conference on Robot Learning*, pages 1691–1729, 2025.

[27] 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*, 2025.

[28] 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*, 2025.

[29] 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.

[30] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083, 2019.

[31] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14738–14748, 2021.

[32] Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. DART: Noise injection for robust imitation learning. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 143–156, 2017.

[33] Chenhao Li, Andreas Krause, and Marco Hutter. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*, 2025.

[34] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of The 8th Conference on Robot Learning*, pages 3705–3728, 2025.

[35] 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.

[36] 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.

[37] Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. In *Robotics: Science and Systems (RSS)*, 2023.

[38] Xiaokang Liu, Zechen Bai, Hai Ci, Kevin Yuchen Ma, and Mike Zheng Shou. World-vla-loop: Closed-loop learning of video world model and vla policy. *arXiv preprint arXiv:2602.06508*, 2026.

[39] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. Serl: A software suite for sample-efficient robotic reinforcement learning, 2024.

[40] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning, 2024.

[41] Ajay Mandlekar, Caelan Garrett, Danfei Xu, and Dieter Fox. Human-in-the-loop task and motion planning for imitation learning. In *Conference on Robot Learning (CoRL)*, volume 229 of *PMLR*, 2023.

[42] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *Robotics: Science and Systems (RSS)*, 2023.

[43] Xiaowen Qiu, Yian Wang, Jiting Cai, Zhehuan Chen, Chunru Lin, Tsun-Hsuan Wang, and Chuang Gan. LuciBot: Automated robot policy learning from generated videos. *arXiv preprint arXiv:2503.09871*, 2025.

[44] 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.

[45] Pengzhen Ren, Kaidong Zhang, Hetao Zheng, Zixuan Li, Yuhang Wen, Fengda Zhu, Mas Ma, and Xiaodan Liang. Surfer: Progressive reasoning with world models for robotic manipulation. *arXiv preprint arXiv:2306.11335*, 2023.

[46] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.

[47] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee,

Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *International Conference on Machine Learning*, pages 30613–30632. PMLR, 2023.

[48] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 30613–30632, 2023.

[49] 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.

[50] Jonathan Spencer, Sanjiban Choudhury, Matt Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. In *Proceedings of Robotics: Science and Systems*, 2020. doi: 10.15607/RSS.2020.XVI.055.

[51] 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.

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

[53] Elise Welte et al. Interactive imitation learning for dexterous robotic manipulation: Challenges and perspectives. *Frontiers in Robotics and AI*, 12, 2025.

[54] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023.

[55] Philipp Wu, Yide Shentu, Qiayuan Liao, Ding Jin, Menglong Guo, Koushil Sreenath, Xingyu Lin, and Pieter Abbeel. RoboCopilot: Human-in-the-loop interactive imitation learning for robot manipulation. *arXiv preprint arXiv:2503.07771*, 2025.

[56] Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-env: Leveraging world model as a virtual environment for vla post-training. *arXiv preprint arXiv:2509.24948*, 2025.

[57] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[58] Kaidong Zhang, Pengzhen Ren, Bingqian Lin, Junfan Lin, Shikui Ma, Hang Xu, and Xiaodan Liang. Pivot-r: Primitive-driven waypoint-aware world model for robotic manipulation. *arXiv preprint arXiv:2410.10394*, 2024.

[59] Zhongru Zhang, Chenghan Yang, Qingzhou Lu, Yanjiang Guo, Jianke Zhang, Yucheng Hu, and Jianyu Chen. Veo-act: How far can frontier video models advance generalizable robot manipulation? *arXiv preprint arXiv:2604.04502*, 2026.

[60] Pengfei Zhou, Liliang Chen, Shengcong Chen, Di Chen, Wenzhi Zhao, Rongjun Jin, Guanghui Ren, and Jianlan Luo. Act2goal: From world model to general goal-conditioned policy. *arXiv preprint arXiv:2512.23541*, 2025.

[61] Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world. *arXiv preprint arXiv:2503.24278*, 2025.

[62] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Mani-WM: An interactive world model for real-robot manipulation, 2024. URL https://openreview.net/forum?id=aVyJwS1fqQ.

V. Appendix

A. Anonymous Interactive Demo

We provide an online interactive demo of Hi-WM at: https://hi-wm.github.io/. The demo allows users to intervene in policy rollouts with keyboard inputs and interact with the learned world model in real time.

B. Acknowledgments

This work is supported by the Sci-Tech Innovation Initiative by the Science and Technology Commission of Shanghai Municipality (24ZR1419000), the National Science Foundation of China (12471501), and ECNU Multifunctional Platform for Innovation (001).

C. Related Work

Post-training with human corrective supervision. Human-in-the-loop robot learning has long studied how to improve policies under the learner-induced state distribution by querying expert correction during deployment or interactive training [37, 26, 55, 41, 53]. More recent work extends this idea to generalist robot policies and post-deployment refinement, where human interventions, preferences, or corrective demonstrations are used to adapt a pretrained policy to downstream tasks [46, 32, 30, 50, 24, 19, 20, 40, 39]. These approaches show the value of targeted supervision near policy failure. However, the corrective loop remains predominantly tied to physical execution: each intervention consumes robot time, scene setup, reset effort, and operator attention in the real world. Our work addresses this bottleneck by moving corrective interaction from the robot to a learned environment.

World models for data generation and policy improvement. A separate line of work uses world models [54, 48, 42] as scalable sources of training signal, including imagination for policy learning, synthetic trajectory generation [31, 5, 9, 4, 3, 62, 47, 45, 58, 33, 22, 27, 52, 2, 49, 17, 56, 29, 12, 38, 16, 25], and world-model-based policy optimization [6, 43, 23, 21, 10, 57, 15, 13, 7, 17, 49, 60, 2, 12, 16, 59]. These methods leverage learned dynamics to expand coverage beyond manually collected demonstrations and reduce dependence on expensive real-world interaction. Despite their differences, the world model is typically used as an **autonomous generator, simulator, or optimizer**. By contrast, our focus is not autonomous rollout alone, but **human-guided corrective interaction inside the world model**, with the resulting trajectories used specifically for post-training around failure-prone states.

World model for policy evaluator. World models are also increasingly studied as evaluation [34, 61, 14] substrates for robot policies, where the central question is whether model rollouts preserve policy ranking, correlate with real-world success, and support tasks such as checkpoint selection or pre-deployment screening [35, 44, 16, 1, 52, 18, 22, 27, 2, 17, 49, 28, 36, 51]. This literature highlights that useful evaluation environments require not only visual realism, but also action faithfulness and consistency with real-world outcomes. Our work is complementary: rather than using the world model only as a post-hoc evaluator, we use it as an interactive corrective workspace in which evaluation, intervention, and corrective data collection are coupled within the same learned environment.

In summary, prior work has explored human corrective learning in the physical world, and world models as generators, evaluators, or optimization environments. Hi-WM lies at the intersection of these directions but addresses a distinct problem setting: **how to use a world model as a reusable workspace for human-in-the-loop post-training**. The defining feature of our formulation is that failure states inside the learned environment can be cached, rewound, and branched into multiple corrective continuations, converting one-shot physical correction into reusable recovery supervision for post-training.

D. Experimental Setup

Real robot tasks. We evaluate our method on three representative manipulation tasks that cover both rigid and deformable object interaction:

- **Fold Towel.** A towel is placed on the table, and the robot is required to grasp two corners and fold it into a target configuration. This task evaluates coordinated manipulation of deformable objects, requiring accurate corner localization and stable execution during folding.
- **Push-T.** A green T-shaped object is placed on a tabletop with a yellow T-shaped target marker. The robot must push and rotate the object to align it with the target pose. This task evaluates precise pose perception and fine-grained planar manipulation.
- **Route Rope.** A rope is randomly placed on the table, and the robot must grasp its two ends, lift them, and place them into two designated slots so that the rope reaches the desired configuration. This task evaluates grasp pose selection and manipulation of flexible objects under large shape variation.

Robot setup. We conduct real-world experiments on a dual-arm YAM-Ultra platform. Each arm provides 6-DoF motion control and a gripper state. Observations are captured from a top-view camera, and robot actions are collected and executed at 15 Hz.

Policy eval initialization. At the beginning of each evaluation rollout, the world model is initialized with a real-world task-start observation, consisting of the initial top-view image of the tabletop scene and the initial poses of the two robot arms. The policy takes the same visual observation and end-effector poses as inputs to generate the first action. The predicted action is then converted into the corresponding end-effector poses and fed to the world model to generate the next observation, which is then used by the policy to produce the subsequent action. In this way, policy evaluation is performed in a closed-loop manner through iterative interaction between the policy and the world model.

E. Efficiency and Scaling of Hi-WM

Cost scaling with virtual intervention data. We first analyze the economic effect of collecting intervention data in the world model rather than in the real world. We define

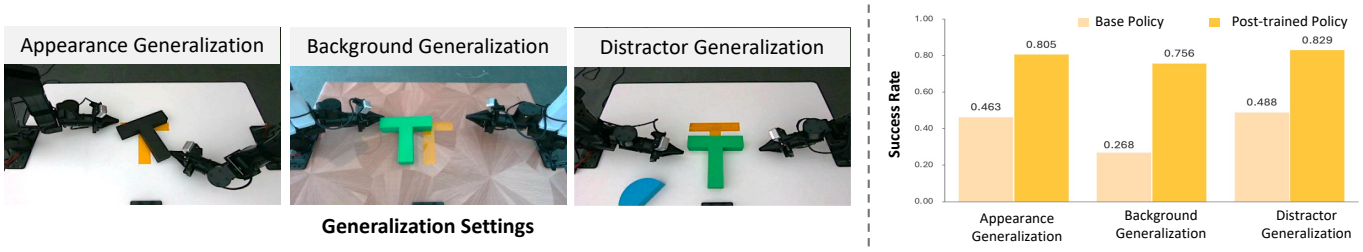

Fig. 5: **Real-world generalization improvements via post-training. Left:** Visualization of diverse real-world generalization settings for Push T, including Appearance Generalization, Background Generalization and Distractor Generalization. **Right:** Quantitative comparison of success rates. The Post-trained Policy, refined using corrective data collected within the Hi-WM, demonstrates substantial performance improvements over the Base Policy across all evaluated generalization scenarios.

the cost saved as $\text{cost}_{\text{real world}} - \text{cost}_{\text{Hi-WM}}$. Our analysis includes both fixed setup costs and time-dependent operating costs, including equipment deployment, human labor, real-world site overhead, and inference compute. As shown in Figure 4(b), the absolute cost savings increase steadily with the number of scenes (N). The gap is smaller at low scene counts and widens as N grows, indicating that Hi-WM has a more favorable marginal cost under increasing scene diversity. Replacing real-world collection with Hi-WM yields nearly $45,000,000 USD in savings. This trend is consistent with the fact that Hi-WM reduces repeated deployment, scene resets, and continuous hardware occupancy during data collection.

Scaling with virtual intervention data. We next study how real-world policy performance changes as more intervention data are collected in the world model. Starting from the same base policy, we add human corrective data collected through Hi-WM in two stages. Specifically, Iteration 1 augments the training set with virtual intervention data amounting to roughly 20% of the original offline dataset size, and Iteration 2 increases this cumulative budget to approximately 35%. We fine-tune the policy after each stage and evaluate the resulting policy in the real world. As shown in Figure 4(c), success rates improve monotonically from the Base Policy to Iteration 1 and again to Iteration 2 across all three tasks. These results indicate that additional corrective data collected in the world model remain beneficial within the tested range and can translate into consistent gains in real-world task success.

F. Policy Generalization via Hi-WM

To answer **RQ4**, we evaluate whether corrective data collected in the world model can improve real-world generalization under held-out scene variations. We focus on Push-T and consider three controlled settings: appearance variation, background variation, and distractor variation. A key advantage of our setting is that the world model can be initialized from varied initial observations under these conditions and rolled out accordingly. This allows human operators to intervene not only on states covered by the original real-world dataset, but also on states induced by these scene variations. We therefore collect corrective data under these varied conditions in the world model, use them for post-training, and evaluate the resulting policy in the real world.

Figure 5 presents qualitative examples of the three settings and a quantitative comparison between the Base Policy and the post-trained policy. The post-trained policy consistently outperforms the Base Policy across all three evaluated settings, indicating that corrective data collected in the world model transfer to improved robustness under these held-out test conditions. These results suggest that Hi-WM improves policy robustness to held-out appearance, background, and distractor variations in the real world.