Wh0: Generative World Models as Scalable Sources of Egocentric Human Hand Manipulation Data

Yangtao Chen[1,2,*], Zixuan Chen[2,*], Peiyang Wang[2,*], Yong-Lu Li[1,3,†], Jing Huo[2,†], Jieqi Shi[2], Yang Gao[2]
[1]Shanghai Innovation Institute, [2]Nanjing University, [3]Shanghai Jiaotong University
*Equal contribution. †Corresponding authors: yonglu_li@sjtu.edu.cn, huojing@nju.edu.cn.

I. Introduction

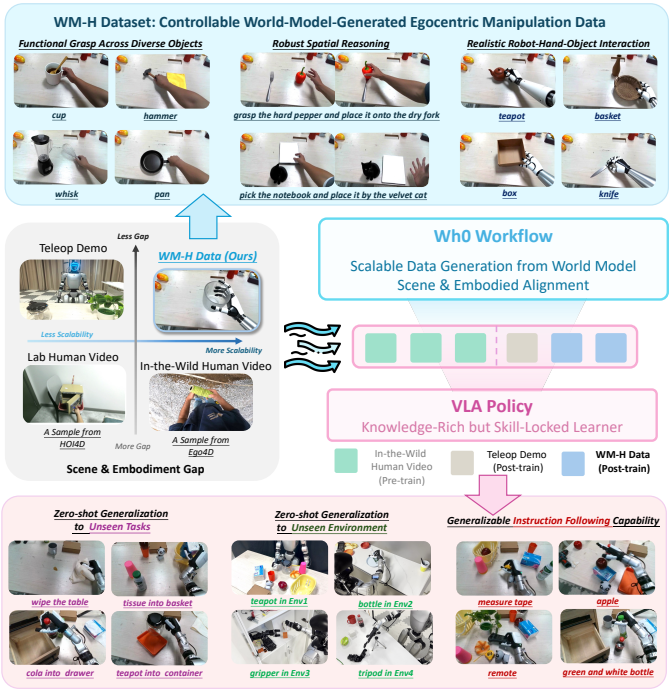


Fig. 1: **Overview of Wh0. Top:** WM-H provides world-model-generated egocentric manipulation videos with diverse objects, layouts, and hand-object interactions. **Middle:** WM-H uniquely combines scale with low scene & embodiment gap to deployment; Wh0 converts them to robot-trainable supervision and co-trains with limited robot data atop a human-video-pretrained VLA. **Bottom:** The resulting policy zero-shot generalizes to unseen tasks, environments, and instructions in real-world manipulation.

Modern VLA models [1], [2], [3], [4] achieve strong generalization by leveraging large-scale data, especially egocentric human videos [5], [6], [7], [8], [9], [10], [11], [12], but still face scene and embodiment misalignment when deployed to dexterous manipulation. Specifically, everyday human environments [13], [14] differ from robot workspaces, leading to a scene gap. Meanwhile, the acting hand in egocentric videos is a human hand rather than a dexterous robot hand [15], [16], resulting in an embodiment gap. Existing data sources address only part of this trade-off. Teleoperation aligns with deployment but is expensive and platform-specific [17], [18]. Simulation scales but suffers from a sim-to-real gap [19], [20]. Real egocentric video scales but remains misaligned [13], [21]. An ideal data source should be scalable and deployment-aligned, while minimizing reliance on human data collection.

In this paper, *we formulate generative video world models as scalable, compute-driven sources for synthesizing egocentric human-hand manipulation data.* Their key advantage lies not in precisely simulating robot dynamics, but in generating diverse human-object interaction videos on demand, conditioned on language, objects, and scenes [22], [23]. Therefore, unlike prior work that uses world models as environment dynamics simulators [24], [25], [26], [27], robot trajectory video generators [28], [29], [30], human-hand generators for downstream retargeting or affordance inference [31], [32], [33], [34], or backbones for training future-predictive policies [35], [36], [37], we focus on their role as controllable human-hand data generators, where scenes, objects, task types, and embodiment appearance become design variables that scale with compute rather than human labor.

As shown in Fig 1, we instantiate this idea with **Wh0**, a framework that uses generative video world models to build egocentric human-hand manipulation data for unlocking dexterous capabilities in pretrained VLA models. Wh0 constructs **WM-H**, a 50k-sample dataset with egocentric manipulation videos, language instructions, and 3D hand motion annotations, via automated instruction generation, scene- and embodiment-aligned video synthesis, visual editing, and hand motion reconstruction. During post-training, Wh0 co-trains WM-H with limited real teleoperated robot data: robot data supplies deployment-specific constraints, while WM-H bridges human manipulation priors and robot execution through a unified hand action space and embodiment alignment.

On a Unitree G1 humanoid with Inspire dexterous hands, under zero-shot evaluation, the strong VLA baseline VITRA [7] post-trained only on robot data without test-task demonstrations achieves an 8.3% success rate, while co-training with WM-H improves it to 38.9%, a 4.7× gain. Ablations verify that scene and embodiment alignment, together with data scale, are the primary drivers of Wh0's gains. A separate analysis shows that WM-H exposes pretraining-acquired manipulation capabilities inaccessible with limited robot data alone.

Our contributions are summarized as follows:

- We formulate generative video world models as scalable, compute-driven sources of egocentric human-hand manipulation data, and construct **WM-H**, a 50k-sample synthetic dataset produced through automated instruction

composition, video generation, visual editing, and hand motion reconstruction.

- We introduce **Wh0**, a framework that converts WM-H into robot-action supervision via cross-embodiment retargeting and synthetic-human/real-robot co-training, adapting pretrained VLA models to dexterous robot deployment.
- We validate Wh0 on 18 real-world dexterous tasks across four scenes, improving zero-shot success on unseen tasks from 8.3% to 38.9%.

II. Related Work

VLA Models and Human-Video Priors. Vision-language-action (VLA) models [1], [2], [3], [4], [38], [39], [40] learn general-purpose policies from large-scale robot and heterogeneous data [17], [41], [42], with recent extensions to dexterous manipulation via hierarchical or VLM-planning frameworks [43], [44]. To reduce reliance on costly real-robot demonstrations, prior work uses egocentric human videos [13], [21], often with hand reconstruction [45], [46], to learn manipulation priors, affordances, and transferable hand actions [47], [48], [49], [50], [15], [6], [5], [7], [51], [8], [52]. However, passive human videos are hard to control and remain misaligned with robot deployment, motivating data sources that are both controllable and deployment-aligned.

World Models for Manipulation. World models serve robotics as dynamics models, visual predictors, and policy substrates [24], [25], [27], and recent video-generation approaches convert synthesized interaction videos into robot supervision such as trajectories, point tracks, or 3D flows [31], [53], [33], [54], [30], [26], [35], with general-purpose engines like Cosmos [23] and GigaWorld [55]. Closest to ours, Large Video Planner [32], Dexterous World Models [34], and DexWM [26] model hand-conditioned dynamics from human videos. In contrast, Wh0 treats generated human-hand videos as controllable, deployment-aligned post-training data, co-trained with limited real robot demonstrations rather than as a dynamics simulator.

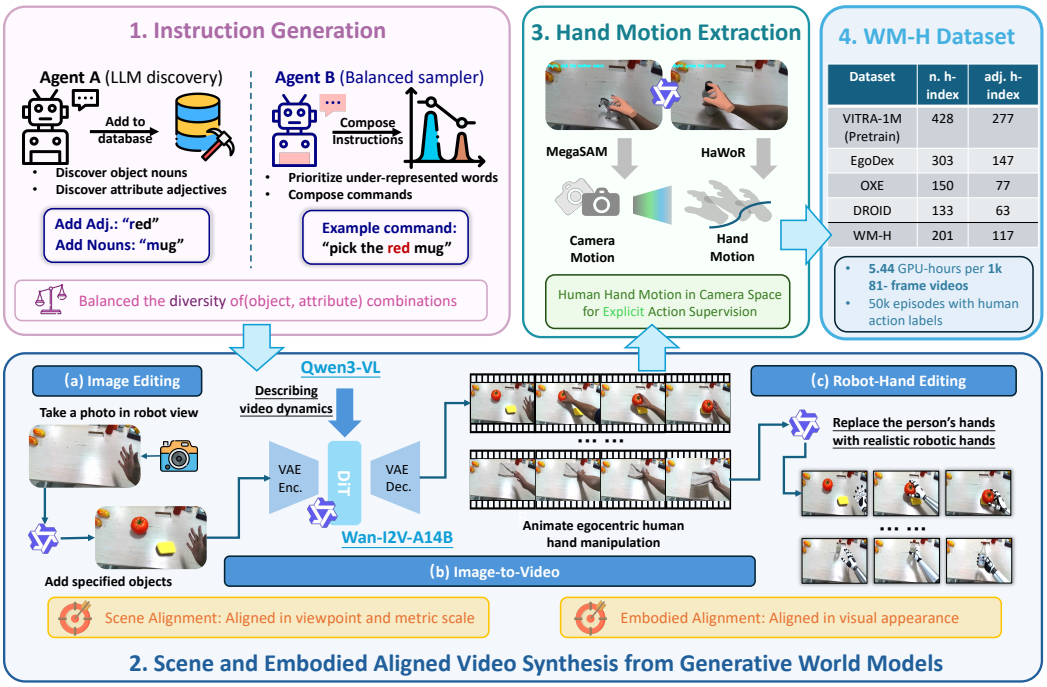
III. WM-H Dataset Construction via Controllable Video Synthesis

We use video world models as controllable generators to produce egocentric human-hand manipulation data. Conditioned on language instructions, object specifications, and scene layouts, the model synthesizes diverse human-object interaction videos. The scalability of the pipeline mainly depends on GPU compute rather than human labor. As shown in Fig. 2, the pipeline consists of instruction generation, scene and embodiment aligned video synthesis, and hand motion extraction, ultimately producing **WM-H**: a 50k-sample egocentric video dataset with 3D hand pose annotations and language task descriptions. The generation cost is approximately 5.44 GPU-hours per 1k videos. We measure instruction diversity using noun and adjective h-indexes [7], where an h-index of h means that at least h words each appear in at least h samples. Despite being much smaller than large-scale human video datasets [7], [6], [17], [18], WM-H achieves broad manipulation-relevant coverage, with a noun h-index of 201 and an adjective h-index of 117 across pick, place, and grasp instructions.

Instruction Generation for Balanced Diversity. A key insight behind our instruction generation pipeline is that a diverse manipulation dataset should contain not only many distinct words, but also sufficient coverage of each word [7]. Simply increasing vocabulary size is insufficient if most words appear only rarely. To balance breadth and frequency, we employ a dual-agent system as shown in Fig. 2(1). The first agent uses a large language model to continually discover object nouns and attribute adjectives related to graspable items, expanding the candidate vocabulary pool. The second agent preferentially samples under-represented words and assembles natural language commands using structured templates such as *pick the* $\{adj\}$ $\{noun\}$. A database tracks word usage frequencies and filters duplicate instructions, enabling systematic and balanced coverage of the space of (object, attribute) combinations.

Video Synthesis with Scene and Embodiment Alignment. Given a generated instruction, we synthesize an egocentric manipulation video through the three-stage pipeline shown in Fig. 2. The pipeline aims to preserve the natural hand-object interaction patterns of video world models while reducing the visual gap to robot deployment: **(a) Scene image capture and editing.** We first capture background images in the target robot workspace for scene alignment. The images are collected by the deployment camera, with the same viewpoint and resolution as the policy input. During capture, we place a human hand in the target interaction region as a scale anchor, providing references for hand size, object scale, and camera-to-workspace distance. Based on the aligned background, we use Qwen-Image-Edit [56] to insert the specified objects into the scene, producing the initial frame for video synthesis. **(b) Image-to-video generation.** We use Wan-I2V-A14B [22] to animate the edited image and generate an egocentric human-hand manipulation video conditioned on the language instruction. To improve the correctness of the generated motion, we use Qwen3-VL [57] to generate a description of the expected hand-object state changes and append it to the video prompt. To enable efficient large-scale synthesis, we adopt LightX2V LoRA adapters, reducing video generation to four inference steps. **(c) Robot-hand editing.** For embodiment alignment, we treat a robotic dexterous hand as a hand entity with a different visual appearance. We use Qwen-Image-Edit [56] to replace the human hand with a realistic robot hand on selected frames, while preserving the original pose, position, object motion, and scene composition. This renders the same manipulation trajectory with both human and robot hand appearances, encouraging the policy to focus on action semantics rather than executor identity.

Motion Extraction for Explicit Action Supervision. Follow [7], we extract 3D hand poses from synthesized human-hand videos as action labels. Generating human rather than robot manipulation videos is important here: human hand reconstruction is comparatively mature, making it feasible to recover reliable motion supervision from generated videos at



Dataset	n. h-index	adj. h-index
VITRA-1M (Pretrain)	428	277
EgoDex	303	147
OXE	150	77
DROID	133	63
WM-H	201	117

Fig. 2: **The overview of the WM-H data synthesis pipeline.** Instruction generation (1) composes balanced (object, attribute) commands; scene- and embodiment-aligned video synthesis (2) edits the deployment scene, animates an egocentric human-hand video, and re-renders the hand as a robot hand; hand motion extraction (3) recovers 3D MANO motion as action supervision.

scale. For each video, we use HaWoR [46] to reconstruct the hand motion. The method first detects hands in each frame and then regresses MANO parameters [58] together with the wrist pose. We keep the wrist pose in camera space and the articulated hand pose in the MANO parameter space. When needed, camera tracking from MegaSAM is used to associate these per-frame predictions with the video camera trajectory.

For a qualitative visualization of WM-H, see Appendix A.

IV. Wh0: Policy Learning with Human-Robot Alignment

a) Policy Architecture: We adopt a VITRA-style vision-language-action policy [7]: a PaliGemma [59] vision-language backbone encodes the current observation and language instruction (with FoV as an auxiliary token for viewpoint cues), and the resulting feature conditions a diffusion-based action decoder that predicts future hand motions from the current hand state.

The action space is defined in the camera coordinate frame of the current observation o_t:

$$a_t = [\Delta t^l, \Delta r^l, \theta_h^l, \Delta t^r, \Delta r^r, \theta_h^r] \in \mathbb{R}^{102}, \tag{1}$$

where $\Delta t, \Delta r \in \mathbb{R}^3$ denote the relative wrist translation and rotation (Euler angles) between consecutive frames, and $\theta_h \in \mathbb{R}^{15\times 3}$ encodes the joint rotations of a 15-DoF MANO [58] hand model in its local frames. Superscripts l, r denote left and right hands; we focus on right-hand manipulation in this work.

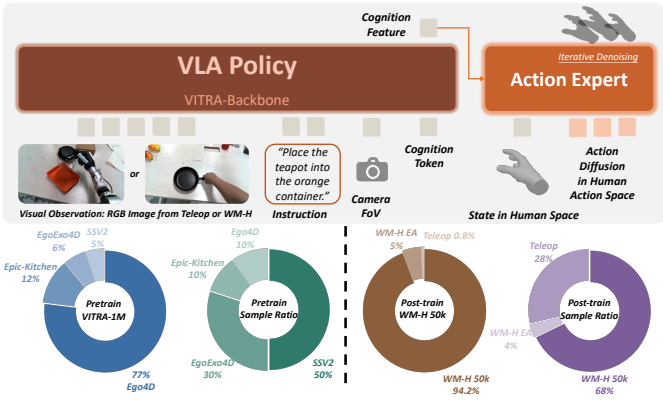


Fig. 3: **Policy architecture and data composition. Top:** A VITRA-style policy denoises actions in the unified MANO space, conditioned on PaliGemma cognition features, FoV, and the current hand state. **Bottom:** Pretraining mixture (VITRA-1M, Ego4D-dominant) and the post-training mixture: Wh0 draws 28% teleop and 68% WM-H (plus 4% embodiment-aligned WM-H), heavily oversampling the scarce robot data given 400 teleop vs. 50k WM-H samples.

We retarget robot joints to MANO and reuse the per-joint normalization parameters precomputed by VITRA [7] from large-scale human videos. This avoids robot-specific normalization from limited data and keeps robot actions aligned with the human action space.

The diffusion action decoder is trained with a standard noise-prediction MSE loss $\mathcal{L}_{\text{MSE}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),\, i}[\,\|\hat{\epsilon}_i - \epsilon\|_2^2\,]$,

Fig. 4: **Real-world evaluation setup.** A Unitree G1 with Inspire hands and a head-mounted egocentric camera (teleoperated via Vision Pro); evaluation spans seen/unseen objects and one seen plus three unseen backgrounds.

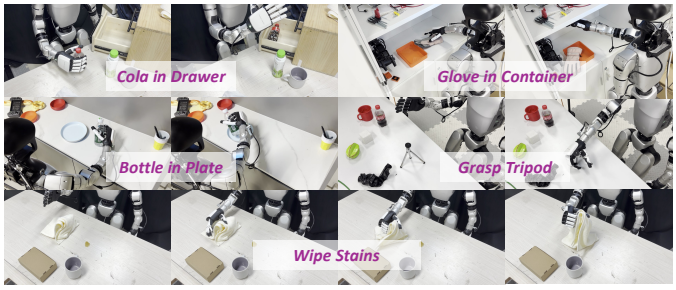


Fig. 5: **Zero-shot rollouts on real-world dexterous tasks**, including container-aware placement, small-object grasping, and tool use. None of these object, container, or task combinations appear in the training set.

where $\hat{\epsilon}_i$ is the predicted noise at diffusion step i.

b) Training Strategy: The policy is initialized from VITRA [7], pretrained on Ego4D [13], Epic-Kitchens [21], EgoExo4D [60], and Something-Something-V2 [61] (Ego4D-dominant mixture), giving the model strong priors on human hand-object interaction. We then co-finetune on a 125:1 mixture of 50k WM-H samples and 400 real teleoperated robot demonstrations. As illustrated in Fig. 3, each batch draws 28% from teleop, 68% from WM-H, and 4% from WM-H EA (WM-H frames after robot-hand editing for embodiment alignment). This oversamples the scarce robot data, providing stable embodiment-specific signals, while WM-H supplies the visual and semantic diversity needed for generalization. We freeze the vision encoder and update the remaining backbone together with the diffusion action decoder, using learning rate 1×10^{-5} and VLM weight decay 0.1.

V. Experiments

Setup. We evaluate Wh0 on a Unitree G1 humanoid with Inspire dexterous hands (Fig. 4); all methods share the same hardware, camera, and control interface. We collect 400 expert trajectories via Apple Vision Pro on seen tasks and backgrounds, mainly pick-and-place over common tabletop objects, and use them as the only real-robot adaptation data. The benchmark contains 18 tasks across four scenes (grasping, placing, and object-specific interactions), each evaluated over 20 trials with randomized object poses; success rate is the primary metric. All policies are evaluated **zero-shot**, without task-specific demonstrations.

Method	Training Setup			Success Rate (%) ↑
	Pretraining	Adaptation Data	Strategy	
$\pi_{0.5}$	Robot	Teleop	FT	$7.78_{\pm 15.6}$
VITRA	Human	Teleop	FT	$8.3_{\pm 8.6}$
VITRA Real Version	Human	Teleop + Real Ego	Co-FT	$21.4_{\pm 23.4}$
Wh0	Human	Teleop + WM-H	Co-FT	$\mathbf{38.9}_{\pm 19.8}$

TABLE I: **Real-world manipulation performance.** We compare different pretraining sources and adaptation data under the same real-robot evaluation protocol. FT denotes fine-tuning on a single adaptation source, while Co-FT denotes joint fine-tuning on multiple data sources.

Baselines and Results. We compare **Wh0** against representative VLA adaptation baselines that differ in pretraining source and adaptation data: $\boldsymbol{\pi}_{0.5}$ [62], a robot-data-pretrained policy adapted with teleoperated demonstrations; **VITRA** [7], a human-video-pretrained VLA adapted with the same teleoperated demonstrations; and **VITRA Real Version**, which replaces WM-H with real egocentric human-hand videos from HOI4D [63] during co-training. As shown in Table I, adapting pretrained policies via robot post-training alone is weak: $\pi_{0.5}$ and VITRA reach only 7.8% and 8.3%, overfitting to seen tasks and losing pretraining generalization. Co-training with additional human data improves zero-shot success, and replacing real egocentric videos with world-model-generated WM-H raises it further from 21.4% to $\mathbf{38.9}\%$, a $4.7\times$ gain over the VITRA baseline. Fig. 5 shows that Wh0 produces more deployment-aligned rollouts on unseen objects, containers, and tasks, indicating that world-model-generated data provides supervision better suited to robot execution. Ablations further attribute these gains primarily to scene alignment, embodiment alignment, and WM-H data scale, while WM-H additionally exposes pretraining-acquired manipulation capabilities that remain inaccessible with limited robot data alone.

For qualitative visualizations of Wh0's real-robot rollouts, see Appendix A.

VI. Conclusion

We presented Wh0, a framework that repurposes generative video world models as scalable, compute-driven sources of egocentric human-hand manipulation data. By synthesizing scene- and embodiment-aligned interaction videos and recovering 3D hand motion as action supervision, Wh0 builds WM-H and co-trains it with limited teleoperated demonstrations to adapt a human-video-pretrained VLA to dexterous robot deployment. On a Unitree G1 with dexterous hands, Wh0 improves zero-shot success on unseen tasks from 8.3% to 38.9%. These results suggest that, beyond acting as dynamics simulators, world models can serve as controllable data engines whose coverage scales with compute rather than human labor. Limitations include reliance on the fidelity of the underlying video generator and a current focus on single-hand pick-place-grasp skills; extending WM-H to bimanual, long-horizon, and contact-rich tasks is a natural next step.

References

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. T. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-1: robotics transformer for real-world control at scale," in *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023* (K. E. Bekris, K. Hauser, S. L. Herbert, and J. Yu, eds.), 2023.

[2] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. T. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han, "RT-2: vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA* (J. Tan, M. Toussaint, and K. Darvish, eds.), Proceedings of Machine Learning Research, pp. 2165–2183, PMLR, 2023.

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning, 6-9 November 2024, Munich, Germany* (P. Agrawal, O. Kroemer, and W. Burgard, eds.), Proceedings of Machine Learning Research, pp. 2679–2713, PMLR, 2024.

[4] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," *CoRR*, vol. abs/2410.24164, 2024.

[5] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A. Cheng, X. Zou, Y. Fang, X. Cheng, R. Qiu, H. Yin, S. Liu, S. Han, Y. Lu, and X. Wang, "Egovla: Learning vision-language-action models from egocentric human videos," *CoRR*, vol. abs/2507.12440, 2025.

[6] R. Hoque, P. Huang, D. J. Yoon, M. Sivapurapu, and J. Zhang, "Egodex: Learning dexterous manipulation from large-scale egocentric video," *CoRR*, vol. abs/2505.11709, 2025.

[7] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, C. Yao, L. Zeng, Z. Feng, H. Liang, S. Xu, Y. Zhang, X. Chen, H. Chen, L. Sun, D. Chen, J. Yang, and B. Guo, "Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos," *arXiv preprint arXiv:2510.21571*, 2025.

[8] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-h0: vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[9] H. Luo, Y. Wang, W. Zhang, S. Zheng, Z. Xi, C. Xu, H. Xu, H. Yuan, C. Zhang, Y. Wang, *et al.*, "Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization," *arXiv preprint arXiv:2601.12993*, 2026.

[10] H. Luo, W. Zhang, Y. Feng, S. Zheng, H. Xu, C. Xu, Z. Xi, Y. Fu, and Z. Lu, "Being-h0. 7: A latent world-action model from egocentric videos," *arXiv preprint arXiv:2605.00078*, 2026.

[11] A. Gavryushin, X. Wang, R. J. Malate, C. Yang, D. Liconti, R. Zurbrügg, R. K. Katzschmann, and M. Pollefeys, "Maple: Encoding dexterous robotic manipulation priors learned from egocentric videos," *arXiv preprint arXiv:2504.06084*, 2025.

[12] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," *arXiv preprint arXiv:2503.00779*, 2025.

[13] K. Grauman, A. Westbury, E. Byrne, V. Cartillier, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, D. Kukreja, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. González, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolár, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbeláez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. A. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, "Ego4d: Around the world in 3,600 hours of egocentric video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 11, pp. 9468–9509, 2025.

[14] Y. Li, X. Wei, J. Luo, Y. Xiao, Y. Bai, G. Zhou, T. Zou, C. Gui, J. Wen, H. Zhang, *et al.*, "Egolive: A large-scale egocentric dataset from real-world human tasks," *arXiv preprint arXiv:2604.23570*, 2026.

[15] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," in *IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025*, pp. 13226–13233, IEEE, 2025.

[16] J. Shi, Z. Zhao, T. Wang, I. Pedroza, A. Luo, J. Wang, J. Ma, and D. Jayaraman, "Zeromimic: Distilling robotic manipulation skills from web videos," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16939–16947, IEEE, 2025.

[17] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. E. Wang, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. P. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Wang, H. Su, H. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Bohg, J. T. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Sünderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante-Gomez, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. D. Sonawani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, L. Xu, X. Li, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, and Z. Lin, "Open x-embodiment: Robotic learning datasets and RT-X models : Open x-embodiment collaboration," in *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*, pp. 6892–6903, IEEE, 2024.

[18] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint*

arXiv:2403.12945, 2024.

[19] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Munoz, X. Yao, R. Zurbrügg, N. Rudin, *et al.*, "Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.

[20] R. Zhao, S. Xu, R. Jin, Y. Deng, Y. Tai, K. Jia, and G. Liu, "Sim2real vla: Zero-shot generalization of synthesized skills to realistic manipulation," in *The Fourteenth International Conference on Learning Representations*, 2026.

[21] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

[22] Wan Team, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[23] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, D. Dworakowski, J. Fan, M. Fenzi, F. Ferroni, S. Fidler, D. Fox, S. Ge, Y. Ge, J. Gu, S. Gururani, E. He, J. Huang, J. S. Huffman, P. Jannaty, J. Jin, S. W. Kim, G. Klár, G. Lam, S. Lan, L. Leal-Taixé, A. Li, Z. Li, C. Lin, T. Lin, H. Ling, M. Liu, X. Liu, A. Luo, Q. Ma, H. Mao, K. Mo, A. Mousavian, S. Nah, S. Niverty, D. Page, D. Paschalidou, Z. Patel, L. Pavao, M. Ramezanali, F. Reda, X. Ren, V. R. N. Sabavat, E. Schmerling, S. Shi, B. Stefaniak, S. Tang, L. Tchapmi, P. Tredak, W. Tseng, J. Varghese, H. Wang, H. Wang, H. Wang, T. Wang, F. Wei, X. Wei, J. Z. Wu, J. Xu, W. Yang, Y. Lin, X. Zeng, Y. Zeng, J. Zhang, Q. Zhang, Y. Zhang, Q. Zhao, and A. Zólkowski, "Cosmos world foundation model platform for physical AI," *CoRR*, vol. abs/2501.03575, 2025.

[24] Y. Matsuo, Y. LeCun, M. Sahani, D. Precup, D. Silver, M. Sugiyama, E. Uchibe, and J. Morimoto, "Deep learning, reinforcement learning, and world models," *Neural Networks*, vol. 152, pp. 267–275, 2022.

[25] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3d-vla: A 3d vision-language-action generative world model," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024* (R. Salakhutdinov, Z. Kolter, K. A. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, eds.), Proceedings of Machine Learning Research, pp. 61229–61245, PMLR / OpenReview.net, 2024.

[26] R. G. Goswami, A. Bar, D. Fan, T. Yang, G. Zhou, P. Krishnamurthy, M. Rabbat, F. Khorrami, and Y. LeCun, "World models can leverage human videos for dexterous manipulation," *CoRR*, vol. abs/2512.13644, 2025.

[27] G. Lu, B. Jia, P. Li, Y. Chen, Z. Wang, Y. Tang, and S. Huang, "Gwm: Towards scalable gaussian world models for robotic manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9263–9274, 2025.

[28] J. Jang, S. Ye, Z. Lin, J. Xiang, J. Bjorck, Y. Fang, F. Hu, S. Huang, K. Kundalia, Y. Lin, L. Magne, A. Mandlekar, A. Narayan, Y. L. Tan, G. Wang, J. Wang, Q. Wang, Y. Xu, X. Zeng, K. Zheng, R. Zheng, M. Liu, L. Zettlemoyer, D. Fox, J. Kautz, S. Reed, Y. Zhu, and L. Fan, "Dreamgen: Unlocking generalization in robot learning through neural trajectories," *CoRR*, vol. abs/2505.12705, 2025.

[29] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 9156–9172, 2023.

[30] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani, "Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation," *arXiv preprint arXiv:2409.16283*, 2024.

[31] J. Liang, R. Liu, E. Ozguroglu, S. Sudhakar, A. Dave, P. Tokmakov, S. Song, and C. Vondrick, "Dreamitate: Real-world visuomotor policy learning via video generation," *arXiv preprint arXiv:2406.16862*, 2024.

[32] B. Chen, T. Zhang, H. Geng, K. Song, C. Zhang, P. Li, W. T. Freeman, J. Malik, P. Abbeel, R. Tedrake, *et al.*, "Large video planner enables generalizable robot control," *arXiv preprint arXiv:2512.15840*, 2025.

[33] H. Li, L. Sun, Y.-H. Hu, D. Ta, J. Barry, G. Konidaris, and J. Fu, "Novaflow: Zero-shot manipulation via actionable flow from generated videos," *arXiv preprint arXiv:2510.08568*, 2025.

[34] B. Kim, T. Kim, J. Lee, and H. Joo, "Dexterous world models," *arXiv preprint arXiv:2512.17907*, 2025.

[35] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian, S. Indupuru, Y. L. Tan, C. Zhu, J. Xiang, *et al.*, "World action models are zero-shot policies," *arXiv preprint arXiv:2602.15922*, 2026.

[36] MotuBrain Team, C. Xiang, F. Bao, H. Liu, H. Tan, H. Bi, J. Li, J. Liu, J. Pang, K. Jing, *et al.*, "Motubrain: An advanced world action model for robot control," *arXiv preprint arXiv:2604.27792*, 2026.

[37] L. Li, Q. Zhang, Y. Luo, S. Yang, R. Wang, F. Han, M. Yu, Z. Gao, N. Xue, X. Zhu, *et al.*, "Causal world modeling for robot control," *arXiv preprint arXiv:2601.21998*, 2026.

[38] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. LLontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu, "GR00T N1: an open foundation model for generalist humanoid robots," *CoRR*, vol. abs/2503.14734, 2025.

[39] Physical Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," in *Proceedings of The 9th Conference on Robot Learning*, 2025.

[40] Physical Intelligence, B. Ai, A. Amin, R. Aniceto, A. Balakrishna, G. Balke, K. Black, G. Bokinsky, S. Cao, T. Charbonnier, *et al.*, "$\pi_{0.7}$: A steerable generalist robotic foundation model with emergent capabilities," *arXiv preprint arXiv:2604.15483*, 2026.

[41] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *conference on Robot Learning*, pp. 991–1002, PMLR, 2022.

[42] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*, pp. 1723–1736, PMLR, 2023.

[43] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang, Y. Yang, and Y. Chen, "Dexgraspvla: A vision-language-action framework towards general dexterous grasping," in *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026* (S. Koenig, C. Jenkins, and M. E. Taylor, eds.), pp. 18836–18844, AAAI Press, 2026.

[44] H. Liu, S. Guo, P. Mai, J. Cao, H. Li, and J. Ma, "Robodexvlm: Visual language model-enabled task planning and motion control for dexterous robot manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2025, Hangzhou, China, October 19-25, 2025*, pp. 1381–1388, IEEE, 2025.

[45] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9826–9836, IEEE, 2024.

[46] J. Zhang, J. Deng, C. Ma, and R. A. Potamias, "Hawor: World-space hand motion reconstruction from egocentric videos," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1805–1815, 2025.

[47] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13778–13790, 2023.

[48] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," in *Conference on Robot Learning*, pp. 654–665, PMLR, 2023.

[49] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *European Conference on Computer Vision*, pp. 570–587, Springer, 2022.

[50] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," in *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024* (D. Kulic, G. Venture, K. E. Bekris, and E. Coronado, eds.), 2024.

[51] R. Zheng, D. Niu, Y. Xie, J. Wang, M. Xu, Y. Jiang, F. Castañeda, F. Hu, Y. L. Tan, L. Fu, *et al.*, "Egoscale: Scaling dexterous manipulation with diverse egocentric human data," *arXiv preprint arXiv:2602.16710*, 2026.

[52] S. Kareer, K. Pertsch, J. Darpinian, J. Hoffman, D. Xu, S. Levine, C. Finn, and S. Nair, "Emergence of human to robot transfer in vision-language-action models," *arXiv preprint arXiv:2512.22414*, 2025.
[53] S. Patel, S. Mohan, H. Mai, U. Jain, S. Lazebnik, and Y. Li, "Robotic manipulation by imitating generated videos without physical demonstrations," *arXiv preprint arXiv:2507.00990*, 2025.
[54] K. Dharmarajan, W. Huang, J. Wu, L. Fei-Fei, and R. Zhang, "Dream2flow: Bridging video generation and open-world manipulation with 3d object flow," *arXiv preprint arXiv:2512.24766*, 2025.
[55] G. Team, A. Ye, B. Wang, C. Ni, G. Huang, G. Zhao, H. Li, J. Zhu, K. Li, M. Xu, Q. Deng, S. Wang, W. Qin, X. Chen, X. Wang, Y. Wang, Y. Cao, Y. Chang, Y. Xu, Y. Ye, Y. Wang, Y. Zhou, Z. Zhang, Z. Dong, and Z. Zhu, "Gigaworld-0: World models as data engine to empower embodied AI," *CoRR*, vol. abs/2511.19861, 2025.
[56] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen, *et al.*, "Qwen-image technical report," *arXiv preprint arXiv:2508.02324*, 2025.
[57] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.
[58] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 245:1–245:17, 2017.
[59] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.
[60] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, E. Byrne, Z. Chavis, *et al.*, "Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
[61] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5842–5850, 2017.
[62] K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. R. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, brian ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," in *9th Annual Conference on Robot Learning*, 2025.
[63] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "HOI4D: A 4d egocentric dataset for category-level human-object interaction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20981–20990, IEEE, 2022.

APPENDIX

Figure 6 provides a qualitative visualization of the WM-H formulation. Figures 7 and 8 present real-robot execution rollouts of Wh0 across a diverse set of manipulation scenarios.

Fig. 6: Qualitative Visualization of WM-H

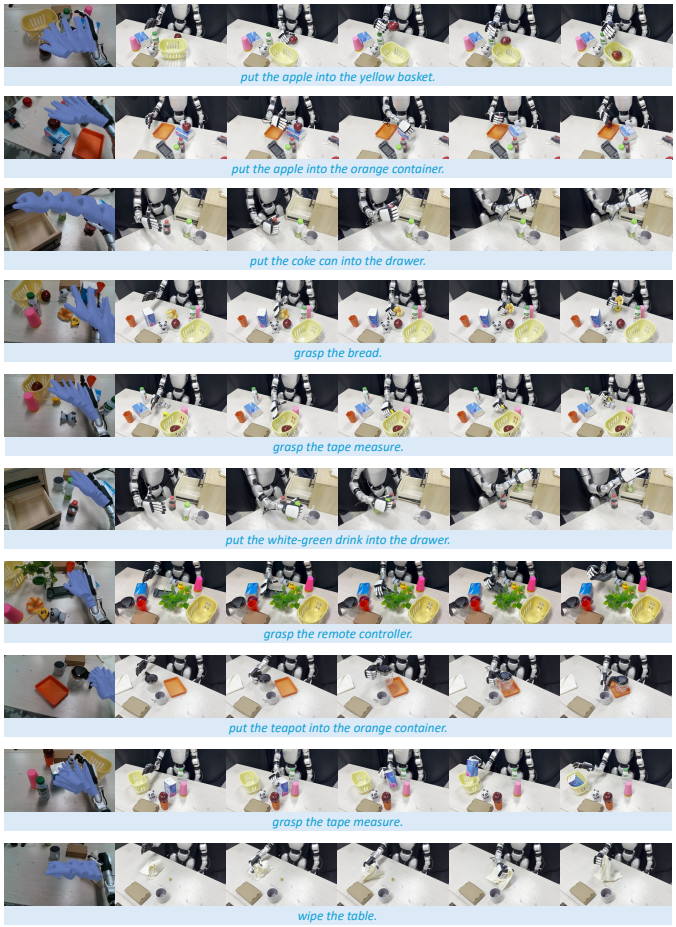


Fig. 7: Robot execution rollouts of Wh0 across various dexterous manipulation tasks.

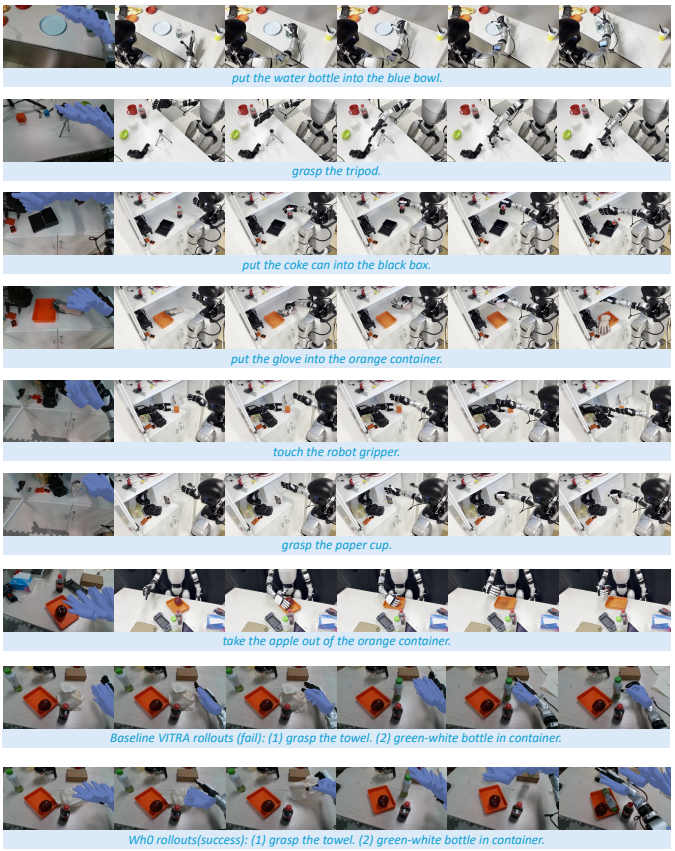


Fig. 8: Additional Wh0 rollouts and comparison with baseline VITRA.