

RoboWM-Bench: A Benchmark for Evaluating World Models in Robotic Manipulation

1. Additional Results for Purely Simulated Robotic Tasks

RoboWM-Bench also includes a set of robotic manipulation tasks evaluated entirely in simulation environments.

The task setup follows the same protocol described in the main paper. For each task, video world models generate future manipulation behaviors conditioned on the simulation observations and the corresponding task descriptions. The predicted behaviors are then converted into executable action sequences and executed in simulation to assess task completion.

Table 1 reports both task-level and step-level execution success rates for these purely simulated tasks. The overall trends are consistent with those observed in the real-to-sim evaluation. As task complexity increases, the success rates of most models decrease, particularly for tasks requiring long-horizon reasoning or precise contact interactions.

Table 1. Task-level and step-level embodied execution success rates (%) on RoboWM-Bench for robotic manipulation tasks evaluated entirely in simulation.

Method		Robot (Task Level)					
		Close Drawer	Push Button	Cut Sausage	Turn Off Faucet	Assemble Burger	Fold Clothes
Cosmos		0%	10%	10%	0%	0%	0%
Wan 2.2		10%	10%	10%	0%	0%	0%
Wan 2.6		30%	20%	40%	0%	0%	0%
Veo		10%	20%	20%	0%	0%	0%

Method		Robot (Step Level)								
		contact	Turn Off Faucet rot. > 20°		turn off	contact	Assemble Burger lift		place	Fold Clothes L. sleeve R. sleeve
Cosmos		10%	0%	0%	10%	0%	0%	0%	0%	0%
Wan 2.2		10%	0%	0%	10%	0%	0%	0%	0%	0%
Wan 2.6		40%	10%	0%	30%	0%	0%	0%	0%	0%
Veo		20%	0%	0%	10%	0%	0%	0%	0%	0%

Table 2. Task-level and step-level embodied execution success rates (%) on the additional human manipulation tasks involving bimanual coordination.

Method	Task Level	Step Level						
	Cook	grasp spatula	grasp pan	lift spatula	lift pan	spatula-pan contact	place spatula	place pan
Cosmos	10%	30%	20%	20%	20%	20%	10%	20%
Wan 2.2	30%	70%	60%	40%	30%	30%	30%	30%
Wan 2.6	50%	90%	70%	70%	70%	60%	50%	60%
Veo	40%	80%	70%	70%	50%	60%	40%	50%
LVP	30%	80%	70%	60%	50%	40%	40%	30%

Method	Task Level	Step Level			
	Lift Large Box	grasp left	grasp right	lift	place
Cosmos	10%	20%	20%	10%	10%
Wan 2.2	30%	70%	60%	30%	30%
Wan 2.6	60%	90%	90%	70%	60%
Veo	50%	80%	80%	50%	50%
LVP	30%	70%	70%	30%	30%

Table 3. **PAI-Bench Domain Score** on Human-Hand Tasks.

Models	Domain Score								Avg.
	Pick Object	Push Button	Put on Plate	Pour Water	Stack Cups	Open Drawer	Put in Drawer	Fold Towel	
Cosmos	96	100	92	72	78	100	100	86	90.5
Wan 2.2	92	100	100	94	100	100	100	100	98.3
Wan 2.6	100	92	100	100	100	100	100	86	97.3
Veo	100	82	100	100	100	92	92	100	95.8
LVP	100	78	100	78	100	100	86	76	87.0

Table 4. **PAI-Bench Domain Score** on Robotic Real Tasks.

Models	Domain Scores								Avg.
	Close Drawer	Pick Object	Push Object	Push Button	Put on Plate	Discard Trash	Pull Object	Put in Drawer	
Cosmos	84	100	100	100	100	100	100	100	98.0
Wan 2.2	100	100	92	100	94	100	98	86	96.3
Wan 2.6	100	100	100	100	100	100	100	100	99.5
Veo	100	100	100	90	100	100	100	80	96.3
Cosmos-FT	100	100	100	100	100	92	100	100	99.0

Table 5. **PAI-Bench Quality Score** on Human-Hand Tasks. Metrics follow the PAI-Bench protocol, including Subject Consistency (SC), Background Consistency (BC), Motion Smoothness (MS), Aesthetic Quality (AQ), Imaging Quality (IQ), Overall Consistency (OC), I2V Subject (IS), I2V Background (IB).

Models	Quality Score								Avg.
	SC	BC	MS	AQ	IQ	OC	IS	IB	
Cosmos	95.0	94.9	99.5	45.8	74.9	25.6	98.3	98.5	79.1
Wan 2.2	95.5	94.9	99.1	41.7	74.5	26.2	98.6	98.5	78.6
Wan 2.6	96.9	95.6	99.2	40.5	75.9	25.8	98.4	97.9	78.8
Veo	96.0	95.5	99.6	40.9	75.4	26.3	98.5	98.6	78.9
LVP	96.3	95.9	99.4	36.4	74.4	25.6	98.1	98.0	78.0

Table 6. **PAI-Bench Quality Score** on Robotic Real Tasks.

Models	Quality Score								Avg.
	SC	BC	MS	AQ	IQ	OC	IS	IB	
Cosmos	94.9	92.9	99.4	44.3	74.5	22.2	94.0	93.5	77.0
Wan 2.2	96.5	94.6	99.5	43.9	73.5	23.2	94.6	93.2	77.4
Wan 2.6	96.9	96.2	99.4	45.1	77.1	23.3	94.3	92.8	78.1
Veo	95.1	92.5	99.6	44.4	70.1	23.6	95.0	94.8	76.9
Cosmos-FT	94.6	94.2	99.5	44.3	64.9	22.8	98.2	98.5	77.1

manipulation task, reporting its average quality score and execution accuracy.

As observed in the scatter plots, most points cluster along a vertical line around an average quality score of approximately 0.78, suggesting that the PAI-Bench quality scores are relatively consistent across different tasks and models. In contrast, the execution accuracy measured by RoboWM-Bench exhibits substantially greater variation. This discrepancy suggests that visual plausibility does not necessarily imply physical correctness, and the embodiment-grounded

evaluation in RoboWM-Bench provides a more sensitive and informative measure of physical executability.

Following the definitions and reporting protocol of PAI-Bench, we additionally present detailed domain and quality scores in tabular form. Note that all scores are computed using videos generated for the manipulation tasks in RoboWM-Bench, rather than the broader video sources used in PAI-Bench. Specifically, the detailed domain scores are reported in Table 3 for human-hand tasks and Table 4 for robotic tasks, while the detailed quality scores are reported

in Table 5 and Table 6, respectively.

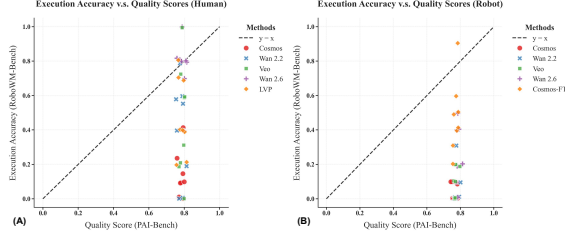


Figure 1. Comparison between the average quality scores in PAI-Bench with the execution accuracy in RoboWM-Bench. The left scatter plot shows human-hand tasks, and the right scatter plot shows robotic tasks.

4. More Visualizations

In this section, we provide additional visualization results for qualitative analysis.

4.1. Predicted Manipulation Videos and Embodied Execution

Figure 4 provides additional qualitative examples of the predicted manipulation videos and their corresponding embodied executions. These results complement the examples shown in Figure 3 of the main paper.

4.2. Real-to-Sim Scene Consistency

Figure 2 provides additional qualitative examples of identical manipulation trajectories executed in real-world scenes and their reconstructed simulation environments, complementing the examples shown in Figure 5 of the main paper.

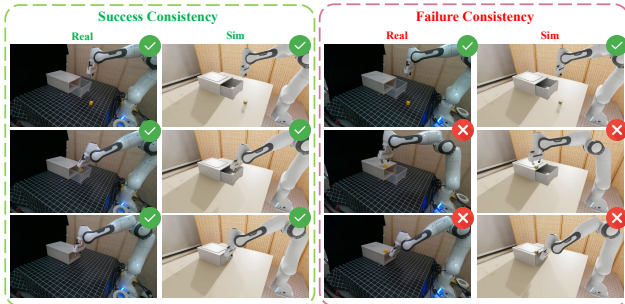


Figure 2. **Additional qualitative results on real-to-sim consistency evaluation.** Identical manipulation trajectories are executed in real-world scenes (left) and reconstructed simulation environments (right). Consistent outcomes demonstrate the fidelity of the reconstruction pipeline.

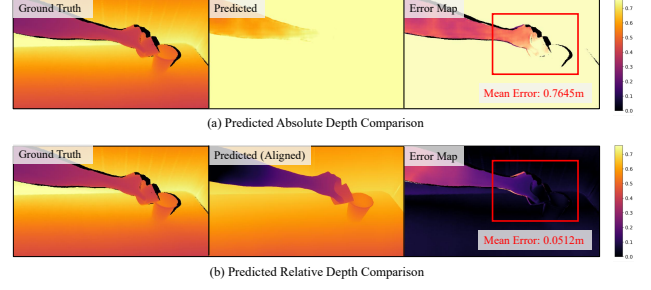


Figure 3. Visualization of depth estimation results. (a) The predicted absolute depth shows a large discrepancy from the ground-truth. (b) Aligning relative depth with the first-frame ground-truth depth improves consistency, although non-negligible errors still remain.

5. Discussion of Depth in Human-Hand Tracking

In our experiments, we empirically found that Phantom [4] provides the most reliable performance for pose tracking and retargeting in human-hand manipulation videos. Therefore, as described in Section 3.2 of the main paper, our pipeline builds upon Phantom with several adaptations.

However, Phantom assumes access to ground-truth depth information, whereas in our setting such depth is not available for the videos predicted by world models. In practice, only the depth of the first frame can be obtained, as it is captured by the camera in the real-world or simulation environment.

Recent works have explored leveraging depth estimation for video understanding [2, 5]. In our setting, we investigate whether estimated depth can improve human-hand tracking. To this end, we estimate depth from RGB frames using Video Depth Anything [3, 6] and evaluate two strategies. First, we directly use the absolute depth predicted by the model. However, as shown in Figure 3(a), the predicted depth shows a large discrepancy from the ground-truth. Second, we estimate relative depth for subsequent frames and align it with the ground-truth depth of the first frame to recover absolute depth values. As illustrated in Figure 3(b), the recovered depth remains imperfect.

Empirically, we observe that incorporating these estimated depths does not improve downstream pose tracking and retargeting performance. Therefore, depth information is not used in the final pipeline.

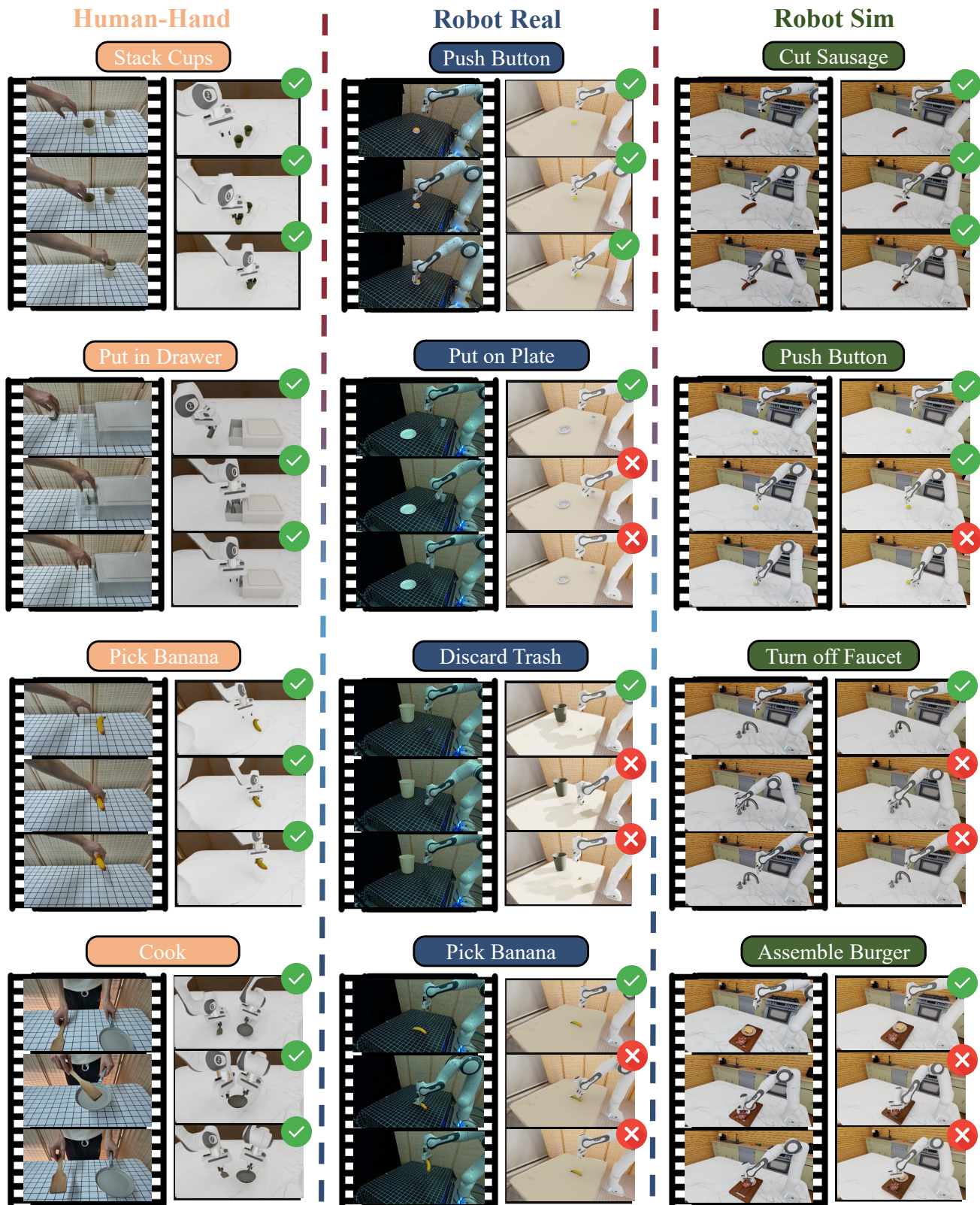


Figure 4. **Additional qualitative results on RoboWM-Bench.** For each task, predicted videos (left) are converted into robot actions and executed in simulation (right).

6. Prompt Details for World Model Video Generation

In this section, we provide details on how instruction prompts are generated for world model video generation.

6.1. Human-Hand Tasks

For human tasks, we employ Qwen3-vl-flash [1] to generate concise descriptions based on initial observations and high-level instructions (e.g., *Pick up the stapler on the table, and stay still.*), as summarized in Table 7 and Table 8. To enhance physical consistency, we append a standardized constraint to the generated instructions: *"The entire human hand must remain fully visible in the frame at all times, with no cropping, no fingers cut off, and no part of the hand outside the camera view. The camera must remain completely static with no movement, no panning, no tilting, no zooming, and no change in viewpoint during the entire video. No extra or unnecessary motions are allowed, the hand must not perform any additional gestures such as turning to show the palm, posing, rotating unnecessarily. All other objects in the scene that are not being manipulated must remain com-*

pletely stationary, with no position shift, no rotation, and no change in placement throughout the entire video." Empirically, these constraints improve the quality of generated videos.

6.2. Robotic Tasks

Similarly, for robotic manipulation tasks, we also leverage Qwen3-vl-flash [1] to generate task-specific descriptions conditioned on initial observations and instructions, as summarized in Table 9 and Table 10. To ensure physical consistency, we append a standardized constraint: *"Throughout the entire video, the gripper undergoes no structural deformation, only the opening angle of its jaws changes; the rotational movements of the robotic arm joints strictly adhere to its inherent mechanical structure; and the camera perspective remains completely unchanged."* Empirically, including these constraints improves the quality of videos generated by the world models.

Table 7. Overview of human-hand manipulation tasks and their corresponding instruction prompts.

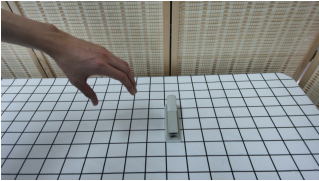
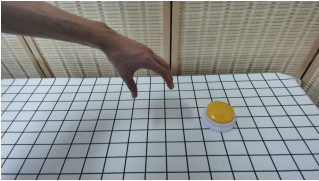
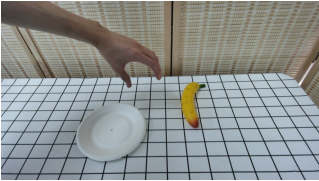
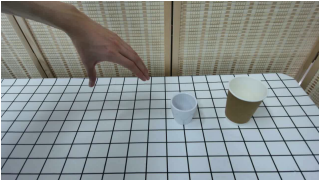
Task Name	Input Image	Task Prompt
Pick Object		The hand lowers slowly toward the stapler and gently grasps it using the thumb and index finger. The stapler is lifted slowly from the table surface without any bouncing or sliding.
Push Button		The hand moves downward toward the button. The fingers make contact with the yellow surface. The hand applies pressure causing the button to compress. The hand lifts up and rest in the air.
Put on Plate		The hand slowly extend downward to gently grasp the banana using the thumb and index finger and then slowly lifts the banana straight up. The banana is then slowly placed on the center of the plate.
Pour Water		The hand grasps the plastic cup using the thumb and the index finger. It slowly lifts the cup from the table. It tilts the cup to pour water into the paper cup. It releases the plastic cup.

Table 7. Overview of human-hand manipulation tasks and their corresponding instruction prompts (continued).

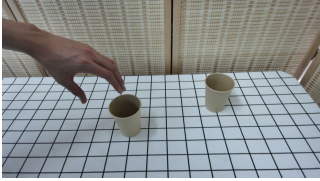
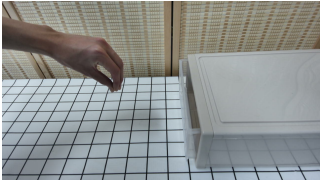
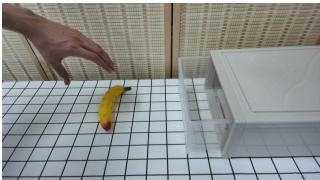
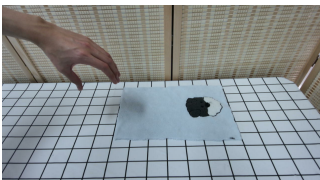
Task Name	Input Image	Task Prompt
Stack Cups		The hand grasps the left edge of the left cup using the thumb and the index fingers and slowly lifts it, then moves it above the right cup and lowers it to stack inside.
Open Drawer		The hand grasps the left edge of the transparent drawer using the thumb and the index fingers then pulls it leftward causing the drawer to slide open.
Put in Drawer		The hand slowly moves to grasp the banana using the thumb and the index finger, lifts it and places it inside the clear drawer. The hand pushes the drawer into the container until it is fully closed.
Fold Towel		The hand grasps the left edge using the thumb and the index fingers and lifts it. The hand folds the towel over to the right edge. The hand aligns the edges and presses down to secure the fold.

Table 8. Overview of human-hand manipulation tasks and their corresponding instruction prompts (continued).


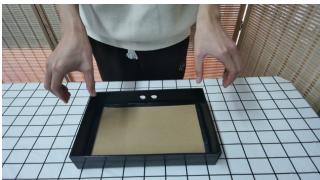
Task Name	Input Image	Task Prompt
Cook		The person uses their left hand to pick up the wooden spatula and places it into the frying pan, while their right hand holds the handle of the pan and lift it up. They move the spatula around inside the pan briefly as if stirring or scraping, then lift the spatula out and place the wooden spatula and the frying pan back on the table in their original position.
Lift Large Box		The person grasps both the left and right edges of the box using the thumb and the index fingers, lifting it vertically off the table.

Table 9. Overview of real-to-sim robotic manipulation tasks and their corresponding instruction prompts.



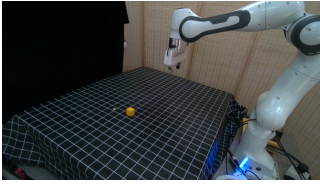










Task Name	Input Image	Task Prompt
Close Drawer		The robotic arm closes its gripper and pushes the drawer back to its closed position.
Pick Object		The robotic arm picks up the white cup from the table surface.
Push Object		The robotic arm closes its gripper and pushes the green cube away from the base of the arm for a short distance.
Push Button		The robotic arm closes its gripper, then uses the gripper to press the yellow button on the tabletop.
Put on Plate		The robotic arm picks up the white cup from the table and moves its gripper above the plate, then releases the cup placing it on the plate.
Discard Trash		The robotic arm picks up the green cube from the table and moves its gripper above the trash bin, then releases the gripper to drop the green cube into the trash bin.
Pull Object		The robotic arm closes its gripper and pulls the green cube toward the robotic arm base for a short distance.
Put in Drawer		The robotic arm picks up the yellow banana, moves it above the open drawer, releases the gripper so the banana falls inside, then closes the gripper to push the drawer back to its closed position.

Table 10. Overview of purely simulated robotic manipulation tasks and their corresponding instruction prompts.

Task Name	Input Image	Task Prompt
Close Drawer		The robotic arm closes its gripper and pushes the drawer back to its closed position.
Push Button		The robotic arm closes its gripper, then uses the gripper to press the yellow button on the tabletop.
Cut Sausage		The robotic arm uses a knife to cut the sausage on the table.
Turn Off Faucet		The robotic arm grasps the lever handle on the left side of the faucet, then rotates it inward to turn off the water.
Assemble Burger		The robotic arm first pushes the meat patty from the left side of the cutting board to its edge, then picks it up and places it on top of the cheese in the plate located on the right side of the cutting board.
Fold Clothes		The left arm grasps the cuff of the left sleeve, while the right arm grasps the cuff of the right sleeve. Both arms then lift and fold the sleeves inward toward the center of the shirt: the left arm places the left cuff onto the left side of the shirt's body, and the right arm places the right cuff onto the right side of the body, aligning them neatly along the torso.

7. Implementation Details of PAI-Bench Domain Score

To facilitate evaluation within the PAI-Bench framework, we design targeted VQA suites for human-hand tasks and robotic tasks, respectively. Within each suite, only the first question is task-dependent and varies across scenarios, while the remaining four questions are identical across all tasks.

7.1. VQA Pairs for Human-Hand Tasks

Question 1: Does the human successfully complete the task: {The task description, such as the human hand grasping and lifting the banana from the table surface.}

options: (A) yes (B) no (C) unclear

ground truth: A

Question 2: Does the human hand make physical contact with the target object?

options: (A) yes (B) no (C) unclear

ground truth: A

Question 3: Does the human hand maintain anatomically plausible hand structure throughout the video (no impossible bends/broken fingers/extra joints)?

options: (A) yes (B) no (C) unclear

ground truth: A

Question 4: Do all task-relevant objects maintain their structural integrity without undergoing physically implausible deformations?

options: (A) yes (B) no (C) unclear

ground truth: A

Question 5: Do any of the task-relevant objects exhibit physically implausible motions, such as sudden spatial displacement?

options: (A) yes (B) no (C) unclear

ground truth: A

7.2. VQA Pairs for Robotic Tasks

Question 1: Does the robot successfully complete the task: {The task description, such as the robotic arm picks up the small yellow cube from the tabletop.}

options: (A) yes (B) no (C) unclear

ground truth: A

Question 2: Does the robot gripper/hand make physical contact with the target object?

options: (A) yes (B) no (C) unclear

ground truth: A

Question 3: Does the robotic system (arm and gripper) maintain structural integrity and a physically plausible configuration throughout the video, with no deformation of rigid links or gripper, and realistic joint rotations?

options: (A) yes (B) no (C) unclear

ground truth: A

Question 4: Do all task-relevant objects maintain their structural integrity without undergoing physically implausible deformations?

options: (A) yes (B) no (C) unclear

ground truth: A

Question 5: Do any of the task-relevant objects exhibit physically implausible motions, such as sudden spatial displacement?

options: (A) yes (B) no (C) unclear

ground truth: A

References

- [1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 5
- [2] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025. 3
- [3] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 3
- [4] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. *URL <https://arxiv.org/abs/2503.00779>*, 2, 2025. 3
- [5] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10486–10496, 2025. 3
- [6] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 3
- [7] Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025. 1