
ThinkAct: Vision-Language-Action Reasoning via Reinforced Visual Latent Planning

—Supplementary Material—

Chi-Pin Huang^{1,2} Yueh-Hua Wu¹ Min-Hung Chen¹
Yu-Chiang Frank Wang^{1,2} Fu-En Yang¹

¹ NVIDIA ² National Taiwan University
{chipinh, krisw, minhungc, frankwang, fredy}@nvidia.com

A Additional Experimental Setup

A.1 Implementation Details

Reinforced Fine-Tuning for Eliciting Visual Latent Planning We set β in GRPO to $1e-2$, with a maximum response length of 1024. To encourage diversity during rollout generation, we set the temperature to 1.0 and use top- p sampling with $p = 0.99$. For computational efficiency, we use up to 16 video frames, each processed at a maximum resolution of $128 \times 28 \times 28$ pixels for video data, and $256 \times 28 \times 28$ pixels for image data. The length of trajectory, K , is set to 8, and for additional QA data, following [5], we use accuracy as the reward for multiple-choice questions, and the average ROUGE-1/2/L scores for free-form answers.

Reasoning-Enhanced Action Adaptation As mentioned in Sec. 4.1, the action model π_ϕ is a Transformer-based diffusion policy [4]. We use a DDPM noise scheduler with 1000 timesteps for training, and inference using 20 DDIM steps. To accelerate training, for each observation o_t and instruction l pair, we let the MLLM \mathcal{F}_θ reason and generate the visual plan latent c_t in an offline manner. With these cached latents, as described in Sec. 3.3, we train the action model π_θ via imitation learning while keeping the VLM frozen. We set the number of interactions per reasoning step N to 15 for SimplerEnv [9] and 75 for the LIBERO benchmark [11], based on the average task length in each environment. We provide an ablation study on the choice of N in Sec. B.6. Following OpenVLA [7], we use a single 224×224 RGB image in third-person view as the observation input during training and inference.

A.2 Training Data Preparation

A.2.1 Training Datasets

2D Trajectory of Manipulation Visual trajectories are sourced from two datasets: Open X-Embodiment (OXE) [16] for robot manipulation, and Something-Something V2 [6] for human manipulation. Specifically, we select the fractal20220817_data and bridge subsets from OXE for their high quality and visually clear trajectories. As described in Sec. 4.1, we extract gripper positions from each frame using an off-the-shelf detector[15]. From each video, we randomly sample 3 starting frames and simplify the subsequent gripper trajectories into K keypoints using the Ramer–Douglas–Peucker (RDP) algorithm (following HAMSTER [10]). For Something-Something V2, we instead use a hand detector [19]. In case two hands appear, we select the one with the largest movement. We apply stabilization [20] to reduce the impact of camera motion.

RoboVQA [18] RoboVQA comprises a diverse set of real-world task episodes collected from both robotic and human embodiments. It contains approximately 5K long-horizon and 92K medium-horizon videos, each annotated with multiple question–answer pairs.

Reflect (RoboFail) [12] The RoboFail dataset captures robot manipulation failures in both simulation and real-world scenarios. It includes 100 simulated failure cases in the AI2THOR environment and 30 real-world cases collected via UR5e teleoperation. We reformulate the original textual annotations into a multiple-choice question format, resulting in a total of 300 question–answer pairs.

EgoPlan-Bench [2] EgoPlan-Bench consists of egocentric videos annotated with task goals, progress histories, and current observations, designed to enhance MLLM planning capabilities in long-horizon daily tasks. It includes EgoPlan-IT, a 50K-instance subset generated automatically, and EgoPlan-Val, a 5K-instance, human-verified subset of high-quality samples.

Video-R1-CoT [5] Video-R1-CoT comprises 165K question–answer samples with chain-of-thought (CoT) annotations generated by Qwen2.5-VL-72B [1]. It is curated to support cold-start fine-tuning for video reasoning and spans domains including math, spatial logic, OCR, and chart understanding. All annotations are filtered for consistency and quality.

LLaVA-Video-178K [21] LLaVA-Video-178K includes 178K videos with detailed captions, 960K open-ended questions, and 196K multiple-choice questions. The annotations are generated via a GPT-4o-based pipeline, providing multi-level temporal descriptions and diverse question types, sourced from untrimmed videos across domains such as cooking, physical activities, and egocentric perspectives.

A.2.2 Training Data Construction

Supervised Fine-Tuning for Cold Start For the SFT cold-start stage, we fine-tune the MLLM using 2D visual trajectories from OXE [16], QA tasks from RoboVQA [18] and EgoPlan-IT [2], as well as chain-of-thought (CoT) data from Video-R1-CoT [5]. Specifically, the SFT dataset comprises 30K 2D visual trajectories, 50K RoboVQA samples, 50K EgoPlan-IT samples, and 165K Video-R1-CoT samples.

For the Video-R1-CoT data, which includes CoT annotations, we follow the original template [5], prompting the model to output responses in the `<think>...</think> <answer>...</answer>` format. For the remaining datasets, which consist of standard QA pairs without intermediate reasoning, we append the instruction: “Please directly provide your text answer within the `<answer> </answer>` tags, without any reasoning process,” to encourage concise responses.

Reinforced Fine-Tuning for Eliciting Visual Latent Planning For the reinforced fine-tuning stage, we use 2D visual trajectories from both OXE [16] and Something-Something V2 [6], along with QA datasets including RoboVQA [18], EgoPlan-IT/Val [2], RoboFail [12], and LLaVA-Video-178K [8]. Specifically, the dataset consists of 12.5K 2D visual trajectories, 10K RoboVQA samples, 10K EgoPlan-IT/Val samples, 0.5K RoboFail samples, and 10K LLaVA-Video-178K samples.

We provide the detailed prompt templates for each data type in Tab. A1. This mixture of action-grounded and reasoning-intensive data enables the model to plan both physically executable and semantically coherent, while also improving generalization to diverse real-world tasks. In our implementation, the format reward r_{format} is set to 1 when the model’s response adheres to the required output structure, specifically the `"<think> ... </think> <answer> ... </answer>"` format, and 0 otherwise.

A.3 Evaluation Benchmarks

SimplerEnv [9] SimplerEnv is a simulation benchmark featuring two evaluation settings: visual matching and variant aggregation. It provides diverse manipulation scenarios across different lighting conditions, table textures, backgrounds, object distractors, and robot camera poses. Built on WidowX and Google Robot setups, SimplerEnv helps assess VLA robustness and the effectiveness of reasoning capability under varied visual conditions.

Table A1: Reasoning prompt template for reinforced fine-tuning.

Data Type	Prompt Template
2D Manipulation Trajectory	<p>Given an image of a robot manipulation scene and the task instruction "{Instruction}", please generate a sequence of 8 keypoints, representing the gripper's 2D trajectory on the image from its current position to the task-completion position. Please think about this planning process as if you were a human carefully reasoning through the manipulation task. Engage in an internal dialogue while considering the scene, the goal, possible subtasks, the motion path, and any obstacles. It's encouraged to include reflections on the environment, analysis of the goal state, decomposition into subtasks, and any adjustments to the planned trajectory as you think through the process. Provide your detailed reasoning between the <think> </think> tags, and then give your final prediction between the <answer> </answer> tags based on the reasoning.</p> <p>Please provide the trajectory [(x1, y1), (x2, y2), ..., (x8, y8)] with coordinates normalized to [0,1] within <answer> </answer> tags.</p>
QA Tasks	<p>{Question} Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags based on the reasoning.</p> <p>(MCQ) Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.</p> <p>OR</p> <p>(Free-form) Please provide your text answer within the <answer> </answer> tags.</p>

LIBERO [11] LIBERO is a simulation benchmark for evaluating generalization in robotic manipulation across four structured task suites, each targeting a distinct generalization challenge: spatial layout variation (LIBERO-Spatial), object diversity (LIBERO-Object), goal variation (LIBERO-Goal), and long-horizon planning with mixed variations (LIBERO-Long). Following prior work [22], we evaluate each task suite over 500 trials using 3 random seeds.

EgoPlan-Bench2 [17] EgoPlan-Bench2 evaluates egocentric planning capabilities of MLLMs in complex, real-world scenarios. It emphasizes long-horizon reasoning based on task goals, progress, and current observations, spanning 24 scenarios across 4 daily-life domains. Compared to EgoPlan-Bench [2], it features more diverse scenes and serves as a non-overlapping evaluation set. The benchmark includes 1,321 high-quality multiple-choice QA pairs evaluated using accuracy.

RoboVQA [18] RoboVQA focuses on visual question answering in robotic manipulation, emphasizing long-horizon reasoning, contextual understanding, and affordance-based decision-making. It includes real-world videos from both robot and human embodiments, covering planning, future prediction, affordance reasoning, and outcome classification. We use its validation set, which consists of 1,893 video-text pairs in a free-form QA format evaluated using BLEU score.

OpenEQA [14] OpenEQA is a benchmark for embodied question answering (EQA), aiming to evaluate an agent's ability to understand and reason about real-world environments through natural language. It poses questions that require spatial, functional, and commonsense understanding across diverse scenes. The dataset includes over 1,600 high-quality human-authored questions from more than 180 real-world environments, in a free-form QA format evaluated using an LLM-based scoring metric aligned with human judgment.

B Additional Experiment Results

B.1 Qualitative Comparisons of Robot Execution Results

To complement the quantitative results, we provide qualitative comparisons of robot execution results between DiT-Policy [4], OpenVLA [7], and ThinkAct, which are available on our project page: <https://jasper0314-huang.github.io/thinkact-vla/>.

B.2 Additional Qualitative Results

Fig. A1 presents a comparison of ThinkAct before and after RL fine-tuning on an EgoPlan-Bench2 [17] example. Similar to Fig. 4 in the main paper, RL enhances embodied reasoning, enabling the model to predict the correct next action.

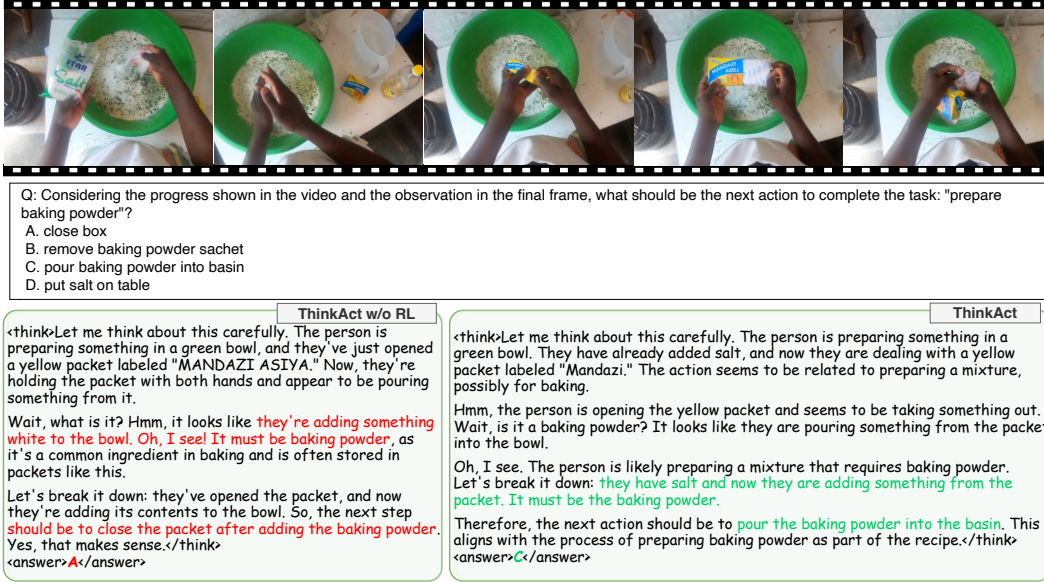


Figure A1: Qualitative comparison of reasoning process and the derived answer for our ThinkAct with and without RL for embodied reasoning tasks on EgoPlan-Bench2 benchmark. **Red** denotes the incorrect reasoning and answer, while **green** indicates the correct one.

B.3 More Self-Correction Samples

To further demonstrate the capacity of ThinkAct for reflection and self-correction, we present two additional examples. In Fig. A2, in a task where the robot is instructed to place a box into a basket, the gripper accidentally drops the box midway. The reasoning MLLM identifies the failure, says "Let's reconsider how to complete the task," and generates a revised plan that guides the gripper back to the dropped location to regrasp the box. The robot then successfully completes the task, demonstrating ThinkAct's ability to reflect on errors and self-correct through structured reasoning.

B.4 Results of Smaller Model Size

To demonstrate the generalizability of our approach, we apply ThinkAct to a smaller model, Qwen2.5-VL-3B, and compare its performance with other models of similar size. As shown in Tab. A2, ThinkAct consistently outperforms other models on EgoPlan-Bench2 [17], RoboVQA [18], and OpenEQA [14], demonstrating its effectiveness on smaller MLLM backbone.

B.5 Results of 5-Shot Adaptation

As shown in Fig. A3, we conduct an additional 5-shot adaptation experiment on LIBERO [11]. Specifically, we fine-tune the action model using only 5 demonstrations per task and evaluate its

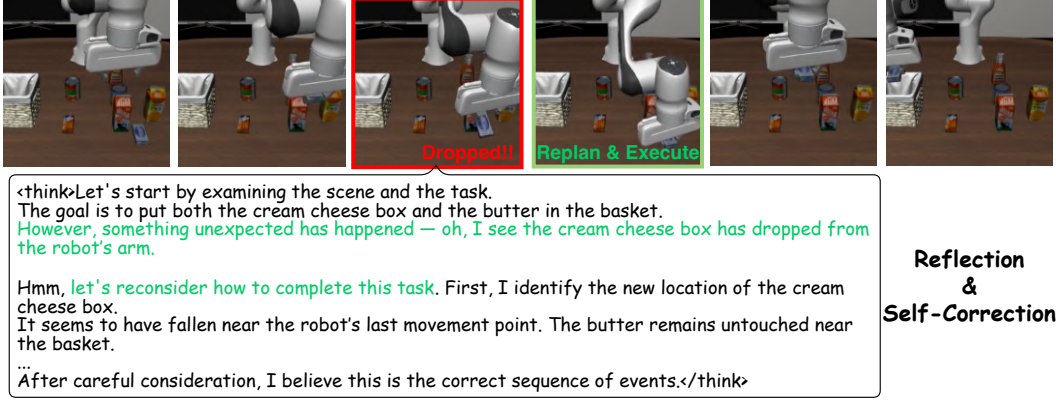


Figure A2: More Demonstrations of self-reflection and correction capability of ThinkAct.

Table A2: Quantitative comparisons with smaller models on embodied reasoning tasks.

Dataset	Split / Metric	InternVL2.5-2B [3]	InternVL3-2B [23]	NVILA-2B [13]	Qwen2.5-VL-3B [1]	Qwen2.5-VL-3B* [1]	ThinkAct-3B (Ours)
EgoPlan-Bench2	Daily life	30.9	36.9	34.6	29.0	44.9	46.6
	Work	27.8	29.9	26.7	27.0	43.0	41.4
	Recreation	28.6	35.6	33.3	30.2	42.2	45.9
	Hobbies	33.1	31.5	31.6	28.9	40.9	42.5
	Overall	30.1	33.4	31.4	28.5	43.0	44.0
RoboVQA	BLEU-1	36.6	34.4	38.7	42.5	60.7	62.4
	BLEU-2	33.7	33.9	34.3	36.3	56.8	57.3
	BLEU-3	31.0	33.5	31.1	28.7	51.3	52.0
	BLEU-4	29.4	33.3	29.2	31.8	45.7	49.6
	Average	32.7	33.8	33.3	34.8	53.6	55.3
OpenEQA	Obj. State	60.5	61.2	59.7	59.8	56.3	60.6
	Obj. Recog.	43.7	42.8	39.6	37.8	41.7	45.3
	Func. Reason.	49.0	53.5	47.2	48.0	45.3	51.4
	Spatial	36.9	38.9	36.5	32.8	36.2	39.4
	Attri. Recog.	63.5	62.6	61.5	57.6	56.6	61.7
	World Know.	42.3	45.2	51.3	38.9	40.9	46.4
	Obj. Loc.	33.6	37.2	33.1	29.0	35.3	37.6
	Overall	47.1	48.8	47.0	43.4	44.6	48.9

performance over 100 trials, following the protocol of Magma [20]. Consistent with the 10-shot results in Fig. 5 of the main paper, ThinkAct consistently outperforms comparative methods across all three tasks.

B.6 Ablation Study

Additional Quantitative Ablation on LIBERO and OpenEQA Benchmarks Tab. A3 extends the main paper’s ablation by evaluating on LIBERO [11] and OpenEQA [14]. Results confirm that both r_{goal} and r_{traj} are crucial for effective planning, with performance dropping when either is removed and nearing the SFT baseline when both are excluded. This further supports the importance of action-aligned visual rewards.

Ablation Study on the Number of Actions per Reason We ablate the frequency of reasoning updates by varying the number of actions per reasoning step N on LIBERO. Setting N to 25, 50, 75, and 100 results in average success rates of 84.0%, 84.6%, 84.4%, and 83.7%, respectively. These results suggest that overly sparse reasoning (e.g., $N=100$) might cause the model to be unable to detect the failure and perform self-correction in time, leading to degraded performance. On the other hand, too frequent updates (e.g., $N=25$) would induce additional inference cost without yielding substantial performance gains. As a result, we set the number of actions per reasoning N as 75 on LIBERO.

Table A3: Quantitative ablation study for our proposed RL rewards in ThinkAct on LIBERO and OpenEQA benchmarks.

Method	LIBERO	OpenEQA
ThinkAct (Ours)	84.4	56.2
Ours w/o r_{traj}	82.1	55.9
Ours w/o r_{goal}	81.7	55.6
Ours w/o $r_{\text{traj}}, r_{\text{goal}}$	81.6	55.7
SFT cold-start	79.1	53.3

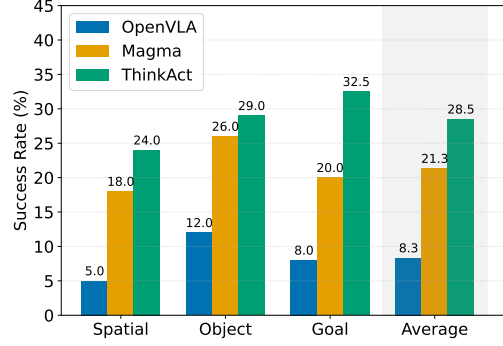


Figure A3: 5-shot adaptation results on LIBERO.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023.
- [3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.
- [5] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- [6] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haefel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.
- [7] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [9] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, 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. *arXiv preprint arXiv:2405.05941*, 2024.
- [10] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, et al. Hamster: Hierarchical action models for open-world robot manipulation. *arXiv preprint arXiv:2502.05485*, 2025.
- [11] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.
- [12] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.
- [13] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024.

- [14] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *CVPR*, 2024.
- [15] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. *arXiv preprint arXiv:2406.11815*, 2024.
- [16] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.
- [17] Lu Qiu, Yuying Ge, Yi Chen, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.
- [18] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024.
- [19] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2020.
- [20] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. *arXiv preprint arXiv:2502.13130*, 2025.
- [21] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.
- [22] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025.
- [23] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.