

# Fast-ThinkAct: Efficient Vision-Language-Action Reasoning via Verbalizable Latent Planning

## Supplementary Material

### A. Limitations

As our verbalizer  $\mathcal{V}_\psi$  is built upon a pre-trained LLM, it inevitably inherits language model limitations including hallucination, occasionally producing plausible but inaccurate descriptions. However, this does not affect action execution during inference, as the verbalizer serves only for interpretability while action prediction uses the grounded latent representations from visual plan distillation. To further improve the faithfulness of verbalized reasoning, we can consider incorporating grounding-aware objectives or hallucination suppression techniques in future work.

### B. Additional Experimental Setup

#### B.1. Algorithm

Algorithm 1 presents the complete training procedure corresponding to Sec. 3.2. It shows how we jointly optimize the teacher model with GRPO and distill its reasoning into the student’s compact latent representations.

#### B.2. Implementation Details

Our implementation follows the setup described in Sec. 4.1 of the main paper. Here we provide additional details. The verbalizer  $\mathcal{V}_\psi$  is initialized from a small LLM, Qwen3-0.6B, with cross-attention layers inserted at each layer to condition on latent CoTs  $\mathbf{z}$ . For the student model training, in the first 3,000 iterations, we replace verbalization loss  $\mathcal{L}_{\text{verb}}$  with language modeling loss using  $\tau^+$  as ground truth to warm up  $\mathcal{V}_\psi$ ’s alignment with the latent representations  $\mathbf{z}$ . We then freeze  $\mathcal{V}_\psi$  and use the  $\mathcal{L}_{\text{verb}}$  for the remaining 1,500 iterations. The student  $\mathcal{F}_\theta$  is optimized throughout both phases. For waypoint prediction in Eq. 6, each  $p_i \in \mathbb{R}^6$  encodes coordinates in the format  $[x_{\text{single}}, y_{\text{single}}, x_{\text{left}}, y_{\text{left}}, x_{\text{right}}, y_{\text{right}}]$ , where the first two dimensions are for single-arm and the last four are for bimanual robot. For ground-truth  $\hat{p}_i$ , we fill the corresponding dimensions based on robot type and mask out the unused dimensions when computing  $\mathcal{L}_{\text{ans}}$ . For GRPO training, we follow the configuration of ThinkAct [9], using rollout size  $N = 5$ . Following [13], we set the number of waypoints in trajectory to  $K = 5$ . We use  $M = 6$  latent reasoning tokens, with ablation study provided in Fig. 11.

During reasoning-enhanced policy learning, for SimplerEnv [14] evaluation, to ensure fair comparison with previous works [12, 13], we initialize  $\pi_\phi$  from DiT-Policy [6] pre-trained on the same OXE dataset [12, 21] and conduct reasoning-enhanced policy learning (Sec. 3.3) using

---

**Algorithm 1:** Training Fast-ThinkAct (Sec. 3.2)

---

**Input:** CoT-SFT checkpoint  $\mathcal{F}_{\theta_0}$ , training data  $\mathcal{D}$ , rollout size  $N$ , latent reasoning steps  $M$ , number of waypoints  $K$ , total iterations  $T_{\text{total}}$

**Output:** Trained student model  $\mathcal{F}_\theta$

```
// Initialize models
1  $\mathcal{F}_\theta^T \leftarrow \mathcal{F}_{\theta_0}, \mathcal{F}_\theta \leftarrow \mathcal{F}_{\theta_0}$ ;
2 Initialize verbalizer  $\mathcal{V}_\psi$  from pre-trained LLM;
3  $t \leftarrow 0$ ;
4 while  $t < T_{\text{total}}$  do
5   Sample batch  $(o, l, \hat{p})$  from  $\mathcal{D}$ ;
   // Suppose bs=1 for simplicity
   // Teacher GRPO training
6   Generate  $N$  rollouts  $\{\tau_i\}_{i=1}^N$  from  $\mathcal{F}_\theta^T(o, l)$ ;
7   Compute trajectory rewards  $\{r_i\}_{i=1}^N$ ;
8   Compute group-wise advantages  $\{A_i\}_{i=1}^N$ ;
9   Update  $\mathcal{F}_\theta^T$  with  $\mathcal{J}_{\text{GRPO}}$  (Eq. 1);
10   $\tau^+ \leftarrow \arg \max_i A_i, \tau^- \leftarrow \arg \min_i A_i$  (Eq. 3);
   // For student distillation
11   $h_t^T \leftarrow$  hidden state of  $\tau^+$  at <answer> token
   from  $\mathcal{F}_\theta^T$ ; // For distillation loss
   // Student latent distillation
12   $\mathbf{z} = \{z_m\}_{m=1}^M \leftarrow \mathcal{F}_\theta(o, l)$ ; // Perform
   auto-regressive latent
   reasoning
13  Compute  $\mathcal{L}_{\text{verb}}$  with  $\mathbf{z}, \mathcal{V}_\psi, \tau^+, \tau^-$  (Eq. 4);
14  Forward  $K$  spatial tokens from  $\mathcal{F}_\theta(o, l, \mathbf{z})$  to
   obtain  $h_t$  and  $\{h'(s_i)\}_{i=1}^K$ ;
15  Compute  $\mathcal{L}_{\text{distill}}$  with  $h_t^T, h_t$  (Eq. 5);
16  Compute  $\mathcal{L}_{\text{ans}}$  with  $\{h'(s_i)\}_{i=1}^K, \hat{p}$  (Eq. 6);
17  Update  $\mathcal{F}_\theta$  with  $\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{verb}} + \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{ans}}$ ;
18   $t \leftarrow t + 1$ ;
19 end
20 return  $\mathcal{F}_\theta$ ;
```

---

the same OXE data. For LIBERO [15] and RoboTwin2.0 [3] evaluations, we initialize  $\pi_\phi$  from RDT [16], which has demonstrated strong performance on RoboTwin2.0, and conduct policy learning using OXE [21] and static ALOHA datasets [24, 28]. Our method further enhances RDT’s manipulation capabilities on both benchmarks. The use of different action models also demonstrates that our approach is agnostic to the underlying action model choice.

## B.3. Training Data Details

### B.3.1. Dataset Sources

**2D Visual Trace of Manipulation Tasks.** For single-arm manipulation, we utilize 2D visual trajectories labeled by MolmoAct [13] from the Open X-Embodiment (OXE) dataset [21], comprising approximately 1.3M trajectories. For bimanual manipulation, we extract dual-arm visual trajectories from the AIST dataset [20], resulting in approximately 92K trajectory samples. Specifically, we first use Molmo-72B [7] to detect left and right gripper positions (following [13]) in the first frame, then apply CoTracker3 [11] to track and parse the manipulation trajectories throughout the video sequences.

**RoboFAC [18].** RoboFAC is a robotic failure analysis dataset containing 9,440 erroneous manipulation trajectories across 16 tasks in both simulated and real-world environments. We utilize the training set with 64K QA pairs covering various failure types for developing failure identification and correction planning capabilities.

**RoboVQA [23].** RoboVQA contains robot manipulation videos with QA tasks covering task understanding. The dataset includes approximately 5K long-horizon and 92K medium-horizon video sequences from diverse robotic platforms, resulting in total 798K QA pairs. Videos are annotated with multiple questions probing spatial reasoning, action prediction, and task comprehension.

**ShareRobot [10].** ShareRobot is a large-scale dataset collected by RoboBrain [10], containing over 1M QA pairs covering task planning, object affordances, and manipulation strategies across diverse robot embodiments and scenes. The dataset features fine-grained annotations linking task descriptions to frame-level execution details, facilitating learning of transferable manipulation knowledge.

**EgoPlan-Bench [4].** EgoPlan-Bench features egocentric videos of daily activities annotated with task planning information including goals, execution history, and current states. The dataset contains approximately 53K video-text pairs for training long-horizon planning and progress tracking capabilities from egocentric view.

**Video-R1-CoT [8].** Video-R1 comprises 165K video question-answer pairs with chain-of-thought reasoning annotations generated by large-scale vision-language models. The dataset covers diverse reasoning domains including mathematical logic, spatial understanding, OCR, and visual analytics. All samples are quality-filtered to ensure annotation consistency and correctness.

**PixMo [7].** PixMo is a general-purpose vision-language dataset with diverse image captions and question-answer pairs. Following MolmoAct [13], we incorporate PixMo dataset to preserve general visual understanding and prevent catastrophic forgetting when training on embodied dataset. Specifically, we use approximately 726K samples from the `ask_model_anything`, `cap`, and `cap-qa` splits.

### B.3.2. Data Processing and Formatting

**Supervised Fine-Tuning (SFT).** To enhance foundational embodied knowledge, we perform supervised fine-tuning on approximately 4M samples combining 2D visual trajectories from MolmoAct [13] and AIST [20], along with QA data from PixMo [7], RoboFAC [18], RoboVQA [23], ShareRobot [10], and EgoPlan [4]. This stage enables the model to acquire basic visual understanding, task comprehension, and manipulation knowledge across diverse embodiments and scenarios.

**Chain-of-Thought SFT (CoT-SFT).** To develop reasoning capabilities while preserving embodied understanding, we sample 5% from the SFT data (approximately 200K samples) and augment with 165K samples from Video-R1-CoT [8]. For data with CoT annotations, we format prompts to elicit structured reasoning enclosed in `<think>` tags followed by answers in `<answer>` tags; for data without CoT annotations, we prompt for direct answers only. This enables the model to learn reasoning capabilities from CoT-annotated data and generalize them to embodied tasks.

**Teacher-Student Training.** Building upon the CoT-SFT checkpoint, we curate a balanced training set by sampling approximately 5,000 instances from each dataset and data type, totaling nearly 50K samples. We adopt the prompt formatting strategy from CoT-SFT for both teacher GRPO training and student latent distillation. We train both the teacher with GRPO and the student with latent distillation (as detailed in Sec. 3.2) on this data, efficiently transferring high-quality reasoning patterns into compact latent representations.

## B.4. Evaluation Setup

### B.4.1. Embodied Reasoning Benchmarks

We evaluate on three benchmarks assessing different aspects of embodied reasoning. EgoPlan-Bench2 [4] tests egocentric task planning across 24 daily-life scenarios with 1,321 multiple-choice questions, measuring accuracy in predicting next steps given task goals and progress history. RoboVQA [23] evaluates visual reasoning in manipulation contexts through 1,893 free-form QA pairs from robot and human demonstrations, assessed via BLEU score. OpenEQA [19] assesses spatial and functional understanding through 1,600+ questions spanning 180+ real-world en-

Table 4. Quantitative results with larger model size (7B or 8B) on embodied reasoning benchmarks.

Method	EgoPlan-Bench2					RoboVQA					OpenEQA	Overall
	Daily.	Work.	Rec.	Hobbies	Avg.	B-1	B-2	B-3	B-4	B-Avg.	Score	Avg.
InternVL2.5-8B [5]	36.2	28.7	34.4	35.4	33.5	40.5	33.3	29.6	27.5	32.7	54.4	40.2
InternVL3-8B [29]	38.5	32.9	36.1	37.2	36.2	44.3	36.5	31.6	28.9	35.3	55.5	42.3
NVILA-8B [17]	35.8	28.7	37.2	35.4	33.7	42.7	39.7	37.6	36.1	39.0	54.0	42.2
Qwen2.5-VL-7B [2]	31.4	26.7	29.5	28.6	29.1	47.8	41.2	36.2	33.7	39.7	50.8	39.9
Magma-8B [26]	32.1	25.7	34.4	29.3	29.8	38.6	31.5	28.1	26.7	31.2	49.1	36.7
RoboBrain2.0-7B [25]	39.4	27.0	33.9	32.2	33.2	44.9	38.2	34.7	33.5	37.8	51.1	40.7
ThinkAct-7B [9]	50.1	49.8	44.8	45.2	48.2	69.1	61.8	56.0	52.4	59.8	56.2	54.7
Fast-ThinkAct-7B	51.3	47.3	41.5	45.9	47.5	70.4	63.3	57.3	53.2	61.1	59.0	<b>55.9</b>

Table 5. Results on LIBERO and SimplerEnv benchmarks with additional ThinkAct-3B comparison.

Method	LIBERO	SimplerEnv-Google	Latency (↓)
OpenVLA-7B [12]	76.5	40.2	N/A
CoT-VLA-7B [27]	83.9	N/A	N/A
ThinkAct-7B [9]	84.4	68.3	7513
MolmoAct-7B [13]	86.8	64.9	6723
ThinkAct-3B [9]	83.1	64.7	5674
<b>Fast-ThinkAct-3B</b>	<b>89.7</b>	<b>68.7</b>	<b>805 (↓7.0×)</b>

Table 6. Comparison with efficient textual reasoning methods.

Method	EgoPlan-Bench2	RoboVQA	OpenEQA	Average
Textual Teacher $\mathcal{F}_\theta^T$	41.7	58.2	49.4	49.8
$\mathcal{F}_\theta^T$ Inference w/o thinking	42.7	55.0	41.7	46.5
$\mathcal{F}_\theta^T$ Inference w/ 6 textual tokens	39.3	53.0	46.5	46.3
$\mathcal{F}_\theta^T$ w/ RL Length-Penalty [1]	41.2	57.5	44.7	47.8
Fast-ThinkAct-3B	<b>46.4</b>	<b>60.8</b>	<b>52.8</b>	<b>53.3</b>

vironments, evaluated using LLM-based scoring aligned with human preferences. These benchmarks comprehensively evaluate embodied reasoning capability across planning, manipulation, and spatial understanding.

#### B.4.2. Robotic Manipulation Benchmarks

We evaluate on three simulation benchmarks covering diverse manipulation scenarios. SimplerEnv [14] provides manipulation tasks with strong sim-to-real correlation, featuring diverse visual variations in lighting, textures, backgrounds, and camera poses. Following MolmoAct [13], we evaluate on the Google Robot tasks using the standard protocol [12, 13] of directly evaluating on SimplerEnv after training on OXE. LIBERO [15] targets different generalization challenges through four task suites: spatial layout variation (LIBERO-Spatial), object diversity (LIBERO-Object), goal variation (LIBERO-Goal), and long-horizon planning with mixed variations (LIBERO-Long). We evaluate each suite over 500 trials using 3 random seeds following prior works [12, 13]. RoboTwin2.0 [3] features

challenging bimanual manipulation with easy and hard difficulty settings, where the hard setting introduces domain randomization including clutter, lighting variations, diverse textures, and height changes. Following the original protocol, we train on 50 clean expert demonstrations per task and evaluate with 100 rollouts under both settings. We assess 10 tasks categorized into short, medium, and long horizons based on demonstration lengths.

## C. Additional Experiment Results

### C.1. Additional Quantitative Results

**Results of Larger Model Size.** To demonstrate the scalability of our approach, we apply Fast-ThinkAct to a larger backbone, Qwen2.5-VL-7B, and evaluate its performance on embodied reasoning benchmarks. As shown in Tab. 4, Fast-ThinkAct consistently achieves strong performance across EgoPlan-Bench2 [22], RoboVQA [23], and OpenEQA [19], validating that our latent reasoning distillation method effectively scales to larger model backbones.

**Performance Comparison with ThinkAct-3B.** Tab. 5 presents detailed numerical results corresponding to Fig. 3 with additional ThinkAct-3B results. At the same 3B model size, Fast-ThinkAct achieves notable performance gains (89.7 vs. 83.1 on LIBERO, 68.7 vs. 64.7 on SimplerEnv-Google) while dramatically improving efficiency with  $7\times$  faster inference (805ms vs. 5674ms).

**Comparison with Efficient Reasoning Baselines.** Table 6 compares our method with efficient textual reasoning alternatives applied to the textual teacher  $\mathcal{F}_\theta^T$ . We evaluate three baselines: removing reasoning during inference entirely (0 tokens), constraining the teacher to generate only 6 textual tokens during inference, and applying RL training with a length penalty [1] to encourage concise reasoning ( $\sim 50$  tokens). These achieve 46.5, 46.3, and 47.8 respectively, all degrading from the teacher’s 49.8. In contrast,

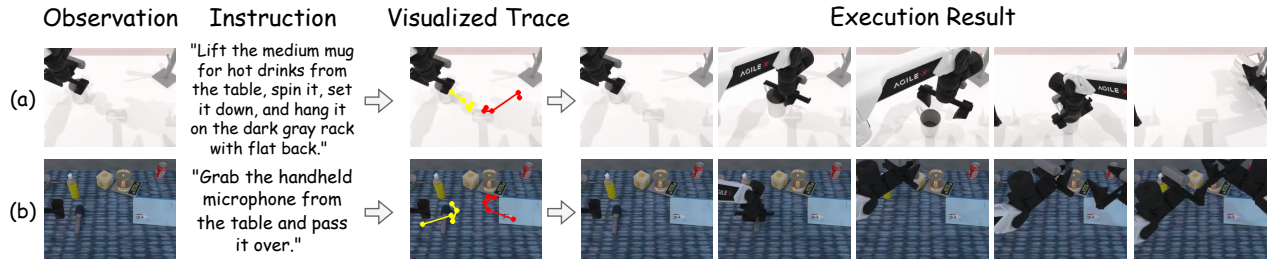


Figure 8. **Visualization of predicted visual trajectories and action execution results on RoboTwin2.0.** Yellow traces indicate left gripper trajectories; red traces indicate right gripper trajectories for bimanual tasks.

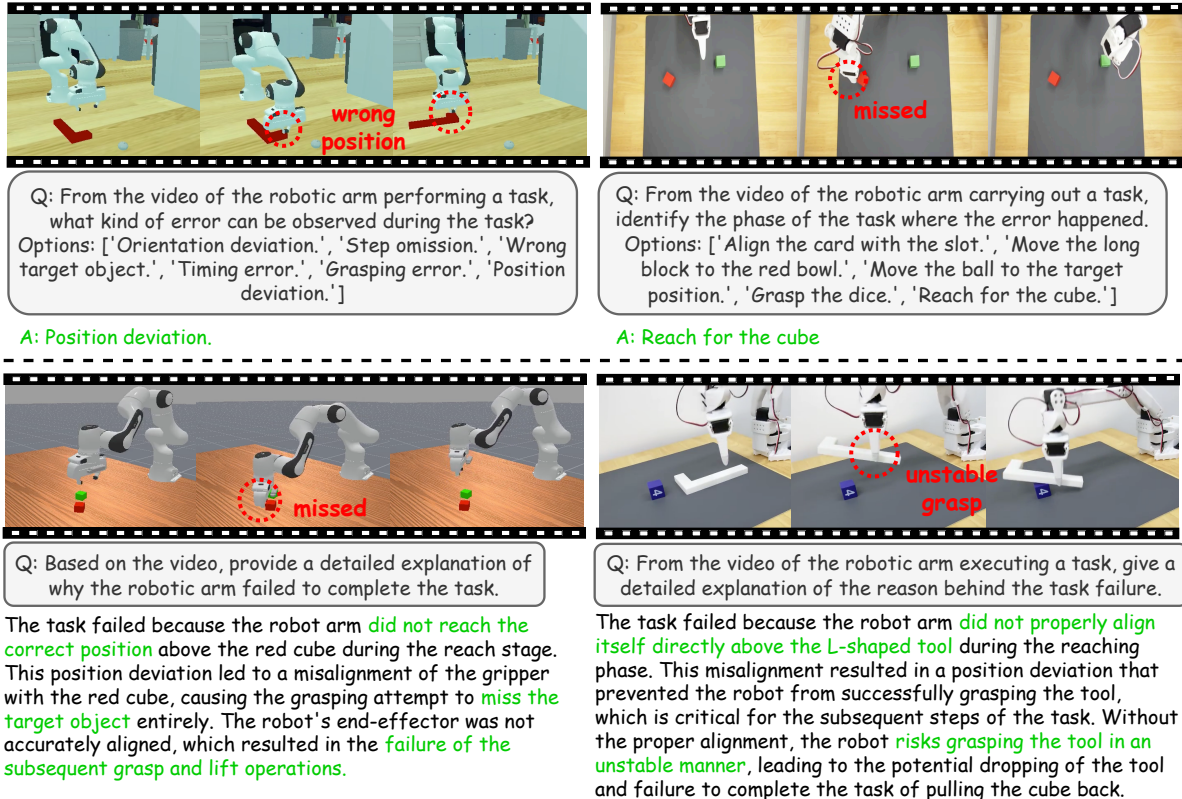


Figure 9. **Failure identification and analysis capabilities on RoboFAC [18].** Top row shows identification of failure types and execution stages. Bottom row demonstrates failure root cause analysis.

Fast-ThinkAct uses only 6 latent tokens and achieves **53.3**, demonstrating superior efficiency and performance.

## C.2. Additional Qualitative Results

**Qualitative Robot Execution.** We provide qualitative robot execution comparisons between the base action model RDT [16] and Fast-ThinkAct in the supplementary video `Fast-ThinkAct.mp4`. Our method shows substantial improvements on challenging robotic execution tasks, where reasoning capabilities provide better spatial understanding and coordination for successful manipulation.

**Bimanual Manipulation Results.** In Fig. 8, we present visualized trajectories and execution results for hanging mug and `handover mic` tasks under easy and hard settings in RoboTwin2.0 [3]. The hard setting includes different backgrounds and distractor objects. These examples show successful bimanual coordination where predicted waypoints accurately guide both grippers through the manipulation sequence, demonstrating Fast-ThinkAct’s spatial reasoning ability across varied visual conditions.

**Failure Identification and Recovery.** In Fig. 9, we demonstrate Fast-ThinkAct’s failure identification and anal-



Question: Is this room on the ground floor of the building?  
GT: no

Let's carefully analyze the image to determine if this room is on the ground floor of the building.

First, I notice that there is a window with a view of greenery outside. **This suggests we are at a level that allows natural light to filter in, which is typical for ground floor windows.**

Next, I see a white desk and a chair, **which seem like standard office furniture for a ground floor workspace.** The desk appears to be at a height that would be comfortable for someone to sit and work, which is also consistent with a ground floor layout.

Additionally, there's a white bookshelf that looks like it could be used for storage and decoration, which is a common feature in office spaces on the ground floor.

However, there is a cabinet in the background that seems a bit unusual for a ground floor room. It could be a storage cabinet for a specific purpose, but it might not be typical for a ground floor space.

Let's also consider the lighting. **The natural light coming through the window seems to be coming from the side, which is more likely if we are on the ground floor** and looking towards a window that is part of the building's exterior.

After considering all these factors, **I think the room is likely on the ground floor,** but it's not impossible for it to be on an upper floor with a unique window design or a specialized storage cabinet.

Final thought: Given the typical layout of office spaces and the visible elements, it's reasonable to conclude that this room is on the ground floor.

(a) Textual reasoning (from teacher model)

Let's break this down. The image shows a study room scene with a desk, a chair, and some shelves. There's no clear indication that this is on the ground floor. It's obviously indoors, but I should check if anything suggests a different level. Wait, I should also consider if there's any sign of elevation. Hmm, it seems like the **view outside the window shows that the room is positioned above the ground.** Oh, I see, the perspective through the window suggests **it's not on the ground floor.** Let's confirm that. Yes, it checks out.

(b) Verbalized reasoning

Figure 10. Reasoning trace comparison on OpenEQA. (a) Teacher's textual reasoning. (b) Student's verbalized latent reasoning. **Green**: reasonable reasoning trace; **red**: incorrect trace.

ysis capabilities, complementing the recovery planning shown in the main paper. The top row shows that Fast-ThinkAct identifies failure types (e.g., position deviation) and execution stages (e.g., reaching for the cube). The bottom row illustrates root cause analysis, for instance, in the bottom-right example, the model correctly infers that the failure to push the cube with an L-shaped tool stems from an improper initial grasp. These results demonstrate Fast-ThinkAct's comprehensive understanding of manipulation failures beyond recovery planning.

**Verbalized Latent Reasoning.** Fig. 10 visualizes teacher textual reasoning and student verbalized reasoning. While the student generates compact and correct (**green**) reasoning, the teacher's lengthy output sometimes contains erroneous steps (**red**) that might degrade the performance.

Table 7. Additional ablation study of training objectives and learning stages on robot manipulation benchmarks.

Method	LIBERO	SimplerEnv Google	RoboTwin2.0	Average
<b>Fast-ThinkAct</b>	<b>89.7</b>	<b>68.7</b>	<b>46.1</b>	<b>68.2</b>
w/o $\mathcal{L}_{verb}$	88.6	67.3	44.9	66.9
w/o $\mathcal{L}_{verb}, \mathcal{L}_{distill}$	86.3	65.7	42.6	64.9
Textual Teacher	88.5	67.3	45.8	67.2
SFT + CoT-SFT	87.2	65.8	43.3	65.4
SFT only	86.9	64.5	42.8	64.7

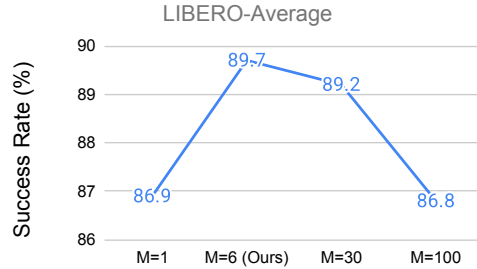


Figure 11. Ablation of Latent Reasoning Steps  $M$ .

### C.3. Additional Ablation Study and Analysis

**Additional Ablation Results on Manipulation Benchmarks.** Table 7 shows ablation results on LIBERO [15], SimplerEnv-Google [14], and RoboTwin2.0 [3]. Removing  $\mathcal{L}_{verb}$  or  $\mathcal{L}_{distill}$  progressively degrades performance, confirming their contributions. Our full model consistently outperforms the textual teacher and models without teacher-student training (CoT-SFT, SFT only), demonstrating the benefits of compact latent reasoning distillation.

**Ablation Study on Action Model Conditioning.** In Sec. 3.3, we extract visual latent planning  $c_t$  from early-layer KV cache of spatial tokens to condition the action model. We compare this against using late-layer KV cache (last  $N$  layers, where  $N$  is the action model depth) and directly using spatial tokens' output hidden states. Our approach achieves **89.7** on LIBERO, outperforming late-layer KV at 88.3 and output hidden states at 87.1, demonstrating that early-layer representations better capture visual planning information for action prediction. Therefore, we adopt early-layer KV conditioning as our default configuration.

**Ablation Study on Latent Reasoning Steps.** In Fig. 11, we study the effect of latent reasoning steps  $M$ . We observe that too few steps ( $M = 1$ ) limit reasoning capacity, while excessive steps ( $M = 30, 100$ ) might introduce redundant or noisy information. Therefore, we adopt  $M = 6$ , which achieves optimal performance, as our default.

## References

- [1] Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025. 3
- [2] 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. 3
- [3] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 1, 3, 4, 5
- [4] 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. 2
- [5] 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. 3
- [6] 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. 1
- [7] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409, 2024. 2
- [8] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 2
- [9] Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025. 1, 3
- [10] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734, 2025. 2
- [11] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025. 2
- [12] 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. 1, 3
- [13] Jason Lee, Jiawei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025. 1, 2, 3
- [14] 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. 1, 3, 5
- [15] 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. 1, 3, 5
- [16] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 1, 4
- [17] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvlla: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024. 3
- [18] Weifeng Lu, Minghao Ye, Zewei Ye, Ruihan Tao, Shuo Yang, and Bo Zhao. Robofac: A comprehensive framework for robotic failure analysis and correction. *arXiv preprint arXiv:2505.12224*, 2025. 2, 4
- [19] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, pages 16488–16498, 2024. 2, 3
- [20] Tomohiro Motoda, Masaki Murooka, Ryoichi Nakajo, Muhammad A. Muttaqien, Koshi Makihara, Hanbit Oh, Keisuke Shirai, Floris Erich, Ryo Hanai, and Yukiyasu Domae. Aist-bimanual manipulation, 2025. 2
- [21] Abby O’Neill, Abdul Rehman, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandelkar, 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. 1, 2
- [22] Lu Qiu, Yi Chen, Yuying Ge, 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. 3
- [23] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debiddatta 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. 2, 3
- [24] Lucy Xiaoyang Shi, Archit Sharma, Tony Z Zhao, and Chelsea Finn. Waypoint-based imitation learning for robotic manipulation. *arXiv preprint arXiv:2307.14326*, 2023. 1
- [25] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 3
- [26] 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. 3
- [27] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 3
- [28] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 1
- [29] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3