ACoT-VLA: Action Chain-of-Thought for Vision-Language-Action Models

Supplementary Material

1. Dataset Description

In this section, we present a comprehensive characterization of the benchmark datasets and the custom-collected data used for model training in our experiments. We systematically report key statistics, including the total number of episodes, frame counts, and other relevant properties, which is summarized in Table 1 below:

Type	Dataset	Embodiment	DoF	Episodes	Frames	FPS
Simulation	LIBERO	Franka	7	1,693	273,465	10
	LIBERO-Plus	Franka	7	14,347	2,238,036	20
	VLABench	Franka	7	4,713	528,398	10
Real-World	Wipe Stain	AgiBot G1	22	177	356,316	30
	Pour Water	AgiBot G1	22	1,821	5,062,506	30
	Open-set Pick	AgiBot G1	22	1,936	219,824	30
	Open-set Pick	AgileX	14	962	251,283	30

Table 1. Dataset statistics.

Simulation Benchmarks. We utilize three publicly released simulation datasets, *i.e.*, LIBERO [3], LIBERO-Plus [1], and VLABench [5]. Specifically, the LIBERO dataset contains $1,693$ episodes and $273,465$ frames, recorded at a fixed 10 Hz. Its demonstrations exhibit relatively uniform trajectory lengths and smooth motion patterns, making it widely adopted benchmark in community.

However, due to the increasing performance saturation observed on LIBERO, LIBERO-Plus is recently introduced to provide a more challenging and diversified evaluation setting. LIBERO-Plus provides $14,347$ episodes and $2,238,036$ frames, captured at 20 Hz. In contrast to the homogeneous trajectories in LIBERO, LIBERO-Plus explicitly emphasizes a perturbation-oriented design. The demonstrations display substantially larger variations in motion magnitude and camera–robot viewpoint configuration. These characteristics make it a more suitable benchmark for evaluating policy generalization under structured distribution shifts.

Besides these two datasets, we further benchmark our method on VLABench, whose training set includes $4,713$ episodes and $528,398$ frames, recorded at 10 Hz, which requires a higher level of visual and physical understanding from the policy.

Real-World Experiment. For real-world deployment, we collect demonstrations across 3 tasks, *i.e.*, Wipe Stain, Pour Water, and Open-set Pick, as shown in Table 1.

The "Wipe Stain" dataset contains 177 episodes with $356,316$ frames, characterized by dense tool–surface contact and fine-grained force control. The "Pour Water" dataset includes $1,821$ episodes and $5,062,506$ frames. Its large scale stems from the task's long-horizon and multi-stage nature. Regarding the "Open-set Pick" task, the AgiBot G1 subset provides $1,936$ episodes with $219,824$ frames, while the AgileX subset offers 962 episodes with $251,283$ frames, both featuring diverse tabletop layouts and natural-language instructions.

Task	Action Space	Action Horizon	State	Batch Size	Training Step
LIBERO	Delta EEF	10	×	128	40K
LIBERO-Plus	Delta EEF	10	×	128	100K
VLABench	Abs EEF	10	✓	128	60K
Wipe Stain	Abs Joint	30	✓	128	50K
Pour Water	Abs Joint	30	✓	128	240K
Open-set Pick	Abs Joint	30	✓	128	50K
Open-set Pick†	Abs Joint	30	✓	128	50K

Table 2. Training details. Note that the "Open-set Pick†" task is performed on AgileX platform.

2. Training & Evaluation Details

Training Details. We describe the task-specific training configurations, *e.g.*, action space and state usage, for better understanding.

As presented in Table 2, for the LIBERO and LIBERO-Plus suites, the policy is trained using delta end-effector control (Delta EEF) with an action horizon of 10 steps. In particular, no privileged state information is provided during training. We utilize a global batch size of 128 and train the policies for 40K and 100K steps, respectively. Similarly, we train our models in VLABench for 60K steps, while adopting state input and absolute end-effector (Abs EEF) actions to align the benchmark's control convention.

In terms of the real-world tasks, we utilize Abs Joint control with a longer action horizon of 30. Unlike the simulator benchmarks, these tasks additionally provide structured robot state observations to improve robustness under real-world sensing and actuation noise. Our models are trained for 50K, 240K, and 50K steps, in "Wipe Stain", "Pour Water", and "Open-set Pick" tasks, respectively, with same batch size of 128.

Evaluation Details. Next, we illustrate the evaluation protocols and success criteria for all real-world tasks. Each task is assessed using fixed and repeatable initializations to ensure reproducibility and reduce environmental variance.

Concretely, in terms of the "Wipe Stain" task, we predefine three initial sponge poses. For each pose, the robot is required to clean stains placed at four distinct table locations. Every configuration is executed twice, resulting in 24 trials in total. A trial is considered successful if the robot grasps the sponge and removes the stain from the specified location.

As for the "Pour Water", we standardize six predefined relative configurations between the bottle and the glass.

Name	EAR	IAR	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
Baseline			70.3	41.7	81.1	97.3	**94.6**	71.8	84.9	75.7
#1	✓		88.7	**63.5**	80.4	94.0	90.2	**89.5**	84.2	83.7
#2		✓	80.7	48.7	**82.6**	**97.7**	90.9	84.3	**86.0**	80.4
#3	✓	✓	**91.2**	62.5	80.3	95.1	91.5	88.3	84.9	**84.1**

Table 3. Module ablations on LIBERO-Plus benchmark. The performance is gradually improved with the addition of proposed methods. Note that models are directly optimized on LIBERO-Plus dataset, with the LLM backbone frozen during training.

Name	Action Head		EAR		LIBERO					LIBERO-Plus							
	Param.	Denoise	Param.	Denoise	Spatial	Object	Goal	Long	Avg.	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
Baseline	300M	10	-	-	98.6	99.0	96.4	92.2	96.6	70.3	41.7	81.1	97.3	**94.6**	71.8	84.9	75.7
#1	600M	10	-	-	97.6	98.4	97.8	96.4	97.6	68.7	44.8	83.1	96.4	92.7	66.6	84.1	74.9
#2	600M	20	-	-	97.8	98.8	98.0	95.2	97.5	70.0	44.8	82.7	**97.6**	93.1	66.7	83.2	75.1
#3	300M	5	300M	5	98.6	**99.6**	97.8	95.4	97.9	88.2	62.4	81.5	95.0	91.5	88.6	**85.3**	**83.9**
#4	300M	10	300M	10	99.0	99.4	98.0	**96.6**	**98.3**	**88.7**	**63.5**	80.4	94.0	90.2	**89.5**	84.2	83.7
#4	300M	10	300M	10	99.0	99.4	98.0	**96.6**	**98.3**	**88.7**	**63.5**	80.4	94.0	90.2	**89.5**	84.2	83.7
#5	300M	10	150M	10	**99.2**	99.2	97.8	94.2	97.6	86.4	54.3	81.7	92.2	91.4	89.1	82.1	81.7
#6	300M	10	250M	10	99.0	98.2	**98.6**	94.2	97.5	87.2	59.7	81.1	95.0	93.7	87.4	83.5	83.1
#7	300M	10	500M	10	98.4	99.4	96.6	94.2	97.0	80.8	57.6	**84.1**	95.6	92.1	79.8	83.7	80.9

Table 4. Effects of parameters and denoise steps on policy performance. Note that the IAR module is not added in this experiment. The evaluation protocol in LIBERO-Plus is *Supervised Fine-Tuning*, *i.e.*, models are directly optimized on LIBERO-Plus dataset. The LLM backbone is frozen during training. The best results are highlighted in **bold**, and the second-best results are underlined.

Then, each configuration is executed two times. A trial is counted as successful if the robot lifts the bottle, pours water into the cup, and places the bottle back onto the coaster. Note that minor spillage of water when pouring is allowed.

Eventually, regarding the "Open-set Pick" task, we initialize ten object arrangements on the table, containing both in-distribution and out-of-distribution instances. In each arrangement, the robot is instructed to grasp a specified target object using either its left or right arm, as indicated by the instruction. Each arm–object pair is evaluated twice, resulting in 40 trials overall. A trial is deemed successful if the robot grasps the instructed object with the correct arm.

Across all tasks, evaluations are carried out by trained operators with substantial prior testing experience, and success rates are computed as the proportion of successful trials relative to the total number of executed attempts.

3. More Experimental Results

In this section, we provide additional quantitative experiments to substantiate the effectiveness of our proposed approach and to empirically uncover several insightful phenomena. Specifically, the experimental analyses comprise four parts: (1) ablation study conducted on the LIBERO-Plus benchmark in Table 3, (2) an investigation of how the parameter sizes of the Action Head and Explicit Action Reasoner (EAR), as well as the number of denoising steps, influence policy performance in Table 4, (3) evaluation of proposed approach's sim2real capability based on Genie-Sim 3.0 [4], and (4) a comparative study examining the relationship among inference latency, model size, and performance in Table 6. Note that we adopt $\pi_{0.5}$ [2] as the baseline method, denoted as "Baseline".

Module Ablation. As shown in Table 3, incorporating the proposed reasoning modules consistently improves policy performance on the LIBERO-Plus benchmark. Adding the EAR module, *i.e.*, experiment '#1', yields a clear gain over the baseline, increasing the average success rate from 75.7% to 83.7%. This improvement can be attributed to EAR's ability to generate an explicit reference action trajectory, which significantly reduces the ambiguity in mapping complex visual or linguistic observations to low-level actions, such as camera shifts and background changes. Meanwhile, incorporating only the IAR ('#2') also improves the performance from 75.7% to 80.4%, indicating that decoding the latent action-related semantics within the vision–language backbone provides useful behavioral priors. Finally, combining EAR and IAR ('#3') achieves the highest success rate of 84.1%, demonstrating their complementary effects, *i.e.*, EAR provides explicit motion guidance, while IAR supplies dense representation-level priors.

Effect of Model Scaling & Denoising Budget. Then, we analyze the superiority of our method by comparing settings with matched total model parameters and denoising steps. As shown in Table 4, firstly, we enlarge the model size of

Tasks	Simulation		Real-World	
	$\pi_{0.5}$	Ours	$\pi_{0.5}$	Ours
Select Color	86.0	**98.8**	85.0	**94.0**
Recognize Size	93.0	**96.0**	**94.0**	**94.0**
Grasp Targets	**71.7**	68.0	70.8	**75.0**
Organize Objects	52.0	**74.0**	60.0	**68.4**
Avg.	75.7	**84.2**	77.5	**82.9**

Table 5. Experimental results on Genie Sim 3.0 Simulation and Real-world Transfer. The best results are highlighted in **bold**.

the action head and increase the number of denoising steps in experiments '#1' and '#2', to construct fair baselines for subsequent comparison. We observe a preliminary observation, *i.e.*, increasing the model size or denoising steps does not reliably enhance performance. Specifically, compared with the baseline, while '#1' improves performance on the LIBERO benchmark, it simultaneously drops on LIBERO-Plus. Next, comparing '#1' and '#2' reveals that further increasing denoising steps yields only negligible fluctuations.

Subsequently, we incorporate the EAR module under fully matched overall parameterization and denoising budgets. Concretely, in both comparison pairs, '#1' with '#3' and '#2' with '#4', we consistently observe notable performance improvements on both benchmarks, once the EAR module is introduced. This indicates that the performance gains originate from our proposed action chain-of-thought. The proposed mechanism supplies explicit reference actions that effectively mitigate the intrinsic instability of action prediction, especially under challenging external perturbations, as shown in the LIBERO-Plus benchmark, enabling more reliable and grounded generalist robotic policy.

Effect of EAR Scale. Moreover, we investigate how various scale of the EAR module influences action prediction fidelity. To isolate the effect of EAR, we keep the action head parameters and the denoising schedule strictly fixed, while scaling the EAR module to 150M, 250M, 300M, and 500M parameters via adjusting hidden size. As presented in Table 4, through the comparison across experiments '#4', '#5', '#6', and '#7', we find that although all EAR-equipped variants outperform non-EAR baselines on both benchmarks, the performance trend is non-monotonic. Applying moderate EAR scales, *e.g.*, 300M, yields the greatest improvement. Particularly, as evidenced in '#7' in Table 4, when the parameter of the EAR module even exceeds that of the action head, we observe a marked drop in performance. We attribute this degradation to the tendency of an over-parameterized EAR to overfit spurious correlations during training. Therefore, it generates reference action trajectories that are systematically biased, which ultimately misdirect the action head toward suboptimal predictions.

Sim To Real Evaluation. To further assess the sim2real transferability of proposed method, we conduct evaluations on the Genie-Sim 3.0 benchmark [4]. It comprises 4 challenging tasks: picking objects based on specified colors, sizes, or categories, as well as a complex tabletop organization task. Technically, the model is trained on the officially released simulation-based datasets. Then, it is deployed and evaluated in both simulation and real-world environments.

Name	EAR	IAR	Param.	Latency	LIBERO Avg. SR	LIBERO-Plus Avg. SR
Baseline			3.35B	91ms	96.9	75.7
#1	✓		3.80B	110ms	98.3	83.7
#2		✓	3.36B	93ms	98.0	80.4
#3	✓	✓	3.81B	112ms	**98.5**	**84.1**

Table 6. Ablation experiment on model efficiency and performance. Note that the evaluation protocol in LIBERO-Plus is *Supervised Fine-Tuning*.

As illustrated in Table 5, our approach consistently outperforms the baseline across both domains. Specifically, it achieves 84.2% in simulation and 82.9% in real-world settings, representing absolute improvements of 8.5% and 5.4%, respectively. Notably, our method exhibits minimal performance degradation during the sim-to-real transition, which we attribute to the fundamental nature of ACoT paradigm. Specifically, while visual domain gap persists between synthetic and real-world observations, the underlying kinematically grounded action guidance remains consistent. By shifting the locus of reasoning from the perceptual space to the action space, our model extracts task-relevant motion priors that are invariant to low-level visual perturbations, effectively enhancing policy's sim2real capability.

Latency Analysis. In Table 6, we further examine the inference efficiency of our approach in terms of both parameter count and end-to-end latency. As additional reasoning modules are introduced, we observe a slight increase in latency. Incorporating the EAR module raises latency from 91ms to 110ms, while adding the IAR module introduces only an additional 2ms. However, this marginal overhead is outweighed by the substantial improvement, which reflects a favorable trade-off.

4. Limitations & Future Works

In this section, we discuss the limitations existing in our work and promising directions for future research.

Although our proposed action chain-of-thought (ACoT) substantially boosts policy performance, our framework still exhibits several constraints. The reasoning modules introduce additional computational cost, which, while relatively modest compared to the performance gains, may pose challenges for deployment on resource-constrained robotic platforms. Besides, another limitation stems from the fact that the prevailing action representation in the community is implemented as action chunks, *i.e.*, sequences of low-level control commands such as joint angles or end-effector poses. While such representations faithfully encode the

executed motions, they lack explicit geometric structure that would facilitate higher-level spatial reasoning, such as object-centric coordination and contact geometry. Hence, the potential of ACoT reasoning may not be fully unleashed. Enriching action representations with spatially grounded information to enable ACoT to operate in geometrically interpretable 3D space, constitutes an interesting and promising avenue for future exploration.

5. LLM Usage Statement

In this paper, we employ Large Language Models (LLMs) solely for minor linguistic refinement during the manuscript preparation stage, such as correcting grammatical errors. None of the technical content, implementation details, or experimental results were generated by LLMs.

References

[1] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. 1

[2] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 2

[3] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 1

[4] Chenghao Yin, Da Huang, Di Yang, Jichao Wang, Nanshu Zhao, Chen Xu, Wenjun Sun, Linjie Hou, Zhijun Li, Junhui Wu, et al. Genie sim 3.0: A high-fidelity comprehensive simulation platform for humanoid robot. *arXiv preprint arXiv:2601.02078*, 2026. 2, 3

[5] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11142–11152, 2025. 1