VLA-R1: Enhancing Reasoning in Vision-Language-Action Models

Angen Ye[1,2]†, Zeyu Zhang[3†], Boyuan Wang[1,2], Xiaofeng Wang[3,4], Dapeng Zhang[1,2], and Zheng Zhu[3*]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] GigaAI, Beijing, China
[4] Tsinghua University, Beijing, China
yeangen2022@ia.ac.cn, zhengzhu@ieee.org

Abstract. Vision-Language-Action (VLA) models aim to unify perception, language understanding, and action generation, offering strong cross-task and cross-scene generalization with broad impact on embodied AI. However, current VLA models often lack explicit step-by-step reasoning, instead emitting final actions without considering affordance constraints or geometric relations. Their post-training pipelines also rarely reinforce reasoning quality, relying primarily on supervised fine-tuning with weak reward design. To address these challenges, we present **VLA-R1**, a reasoning-enhanced VLA that integrates Reinforcement Learning from Verifiable Rewards (RLVR) with Group Relative Policy Optimization (GRPO) to systematically optimize both reasoning and execution. Specifically, we design an RLVR-based post-training strategy with verifiable rewards for region alignment, trajectory consistency, and output formatting, thereby strengthening reasoning robustness and execution accuracy. Moreover, we develop **VLA-CoT-13K**, a high-quality dataset that provides chain-of-thought supervision explicitly aligned with affordance and trajectory annotations. Furthermore, extensive evaluations on in-domain, out-of-domain, simulation, and real-robot platforms demonstrate that VLA-R1 achieves superior generalization and real-world performance compared to prior VLA methods. We plan to release the model, code, and dataset following the publication of this work.

Keywords: VLA models · Chain-of-Thought · RLVR

1 Introduction

Vision–Language–Action (VLA) models unify perception, language, and action. They first learn open-vocabulary semantics and cross-modal alignment from internet-scale image–text pretraining. These semantics are then grounded into the action space through multi-task manipulation data. This enables analogical

†Equal contribution. *Corresponding author.

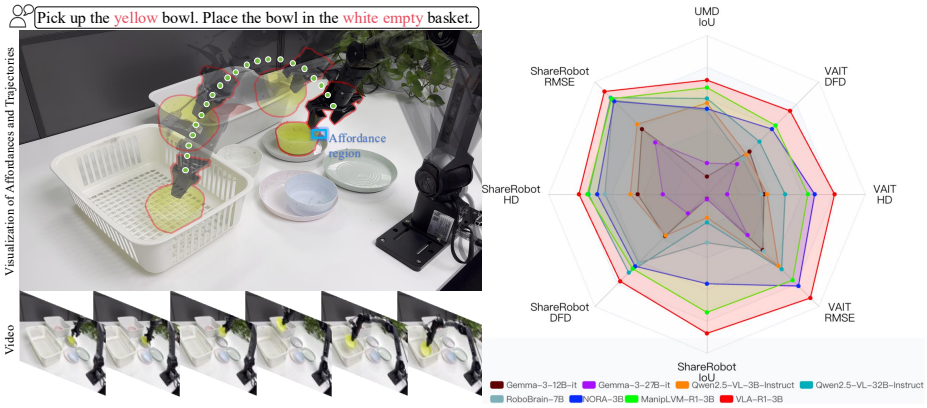

Fig. 1: VLA-R1: pipeline from instruction to execution, with benchmark comparisons against baselines.

transfer to unseen objects and compositional generalization to novel commands. Compared with modular pipelines [9,44] or state-driven policies [11], VLAs show stronger cross-task and cross-scene generalization [18–20, 28, 36, 40, 43]. Representative works include VoxPoser [15] for zero-shot trajectory planning, and ManipLVM-R1 [36] and RoboBrain [19] for integrating affordance perception and pose estimation. Meanwhile, Reinforcement Learning from Verifiable Rewards (RLVR) enhances reasoning and generalization in vision–language models. Vision-R1 [16] matches larger models through cold-start data and progressive training; LMM-R1 [32] employs a two-stage regimen from textual reasoning to multimodal tasks; and VLM-R1 [35] applies R1-style reinforcement to visual grounding, boosting open-vocabulary detection.

However, existing VLA models present two significant challenges. First, they often lack step-by-step reasoning: models tend to emit final actions directly without explicit inference over affordance constraints, geometric relations, or container selection. This limitation leads to instruction-disambiguation failures under color similarity, duplicate instances, or multiple candidate receptacles. Second, post-training rarely provides systematic reinforcement of reasoning. Current method relies on supervised fine-tuning (SFT) with little reward optimization targeted at reasoning quality and execution efficacy. Even when Reinforcement Learning (RL) is used, reward design is typically single-objective and struggles to jointly optimize region alignment and trajectory consistency, degrading performance on out-of-distribution data and in the real world.

To address these challenges, we propose **VLA-R1**, a post-training-enhanced VLA model capable of step-by-step reasoning. Unlike prior approaches, VLA-R1 simultaneously emphasizes data-level Chain-of-Thought (CoT) supervision and optimization-level reward alignment, bridging the gap between reasoning and execution. This enables the model to not only provide answers but also explain

Table 1: Comparison of datasets on affordance, trajectory, reasoning, scenes, and robots

Dataset	#Aff	#Traj	#Reasoning	#Scenes	#Robot
UMD	✓	✗	✗	3	–
VAIT	✗	✓	✗	–	13
VLA-IT	✗	✓	✓	24+	2
ShareRobot	✓	✓	✗	102	12
VLA-CoT-13K	✓	✓	✓	102	12

them, making it robust to challenges like color similarity, repeated instances, and multiple receptacle choices during reasoning.

To further enhance the model's reasoning capabilities, we introduce an RLVR-based post-training strategy at the optimization layer. Specifically, we employ Group Relative Policy Optimization (GRPO) [34] with three verifiable rewards: an affordance reward based on Generalized Intersection over Union (GIoU) [33] to provide informative gradients for non-overlapping predicted and ground truth affordance regions, speeding up learning; a distance-based reward using the improved Fréchet distance to ensure reasonable trajectory curvature and segment length; and an output-format reward to enforce well-formed reasoning and action specifications. These optimizations enable VLA-R1 to generate accurate affordance regions and well-formed execution trajectories, enhancing decision-making.

Moreover, many existing datasets, although large in scale, fail to fully support complex reasoning tasks due to the lack of detailed explanations and reasoning processes in their annotations. To address this, we develop the VLA-CoT data engine, which generates the high-quality **VLA-CoT-13K dataset**, making reasoning steps explicit. The engine aligns CoT with affordance and trajectory annotations, encouraging the model to learn task-consistent reasoning and enabling it to acquire basic reasoning capabilities during the SFT phase.

Finally, we conduct comprehensive evaluations of VLA-R1 across in-domain, out-of-domain, simulation, and real-robot settings. Empirically, VLA-R1 achieves an IoU of 36.51 on the in-domain affordance benchmark, a 17.78% improvement over the baseline; on the in-domain trajectory benchmark it attains an Average distance of 91.74 (lower is better), reducing the baseline by 17.25%. It also delivers state-of-the-art (SOTA) performance in the out-of-domain setting. On physical hardware, VLA-R1 reaches 62.5% success for affordance perception and 75% for trajectory execution. These results demonstrate the method's effectiveness under controlled conditions and its robustness and practicality across domains and real-world scenarios.

Contributions in our paper can be summarized in the following three folds:

– We propose **VLA-R1**, a VLA foundation model that introduces an RLVR optimization scheme with carefully designed rewards for region alignment, trajectory consistency, and output formatting. The optimization is further

enhanced by GRPO, which systematically strengthens reasoning and execution robustness while reducing reliance on manual annotation.
- We introduce the VLA-CoT data engine, which produces high-quality **VLA-CoT-13K** aligned with affordance and trajectory labels and incorporates verifiable rewards, explicitly remedying the lack of step-wise reasoning in existing VLA models.
- Experiments show that VLA-R1 achieves **state-of-the-art** performance across benchmarks. Compared with the strongest baseline, VLA-R1 improves affordance localization by approximately 17.8% while reducing overall trajectory errors by about 17.3%. In real-world evaluations, VLA-R1 achieves the highest task success rates across scenarios, reaching 62.5% for affordance perception and 75% for trajectory execution, demonstrating the effectiveness of our training paradigm

2 Related Work

2.1 VLA Models

Early manipulation research often relies on state-based RL [2, 10], but these methods struggle with high-dimensional visual inputs. More recently, vision-centric approaches have become dominant, harnessing the reasoning capabilities of large language models (LLMs) to improve generalization [7,23,27,39,42]. VoxPoser [15] uses vision-language models to generate 3D value maps, enabling zero-shot trajectory planning. RoboFlamingo [24] fine-tunes on manipulation datasets to perform language-conditioned tasks, while ManipLLM [23] incorporates CoT reasoning to integrate object understanding, affordance perception, and pose prediction into an interpretable framework. Building on this line, OpenVLA [20] and RoboMamba [27] leverage fine-grained CoT data and SFT for further performance gains. Other works, such as Embodied-Reasoner [46], Cosmos-Reason1 [3], and RoboBrain [19], focus on long-horizon reasoning, interpretability, and logical consistency in manipulation tasks. Despite progress, most approaches still depend on large-scale annotated datasets. ManipLVM-R1 [36] reduces reliance on supervision by combining small amounts of labeled data with RLVR-based self-improvement, yielding robust generalization under limited supervision. In contrast, our VLA-R1 explicitly integrates chain-of-thought supervision with RLVR-based optimization, closing the gap between reasoning quality and action execution.

2.2 RLHF for MLLMs

Multimodal Large Language Models (MLLMs) have demonstrated remarkable reasoning capabilities across diverse visual tasks [4, 13, 14, 22, 25, 41]. Recently, RLVR has emerged as a promising way to enhance their reasoning abilities. For example, Vision-R1 [16] leverages a cold-start math dataset and progressive Thinking Suppression Training to achieve results comparable to much larger

models without relying on human annotations. LMM-R1 [32] adopts a two-stage framework, first refining reasoning on textual data and then extending to multimodal and agent-based reasoning tasks. Similarly, VLM-R1 [35] applies an R1-style RL approach to visual grounding, improving open-vocabulary detection and generalization. While these works highlight RLVR's potential, their scope remains limited to non-embodied domains. To bridge this gap, ManipLVM-R1 [36], adapts RLVR to robotic manipulation, enhancing both reasoning and action execution.

3 Method

3.1 Overview

The overall architecture of **VLA-R1** is shown in Fig. 2. Given an input image and a natural language instruction, VLA-R1 encodes multimodal information with a vision–language backbone and produces a structured prediction that is converted into executable robot commands by a deterministic control stack. The vision branch processes the raw image with a visual encoder and projects features into a shared embedding space, while the language branch tokenizes and embeds the instruction. The multimodal decoder fuses the two modalities and jointly reasons over visual cues, textual context, and temporal history, generating an output composed of an explicit reasoning trace and geometric action specifications. These specifications include a 2D interaction region, represented as an affordance box, and a 2D execution trajectory. At inference time, an action decoder parses the geometric predictions, performs depth back-projection using calibrated camera parameters, and applies robot kinematics to obtain joint-level commands for execution. This design bridges high-level task descriptions with grounded low-level control while preserving interpretability through explicit reasoning traces.

3.2 Embodied CoT Data Synthesis

To explicitly inject step-by-step reasoning into vision-language-action learning, we augment the raw ShareRobot-style supervision with chain-of-thought annotations generated by a strong teacher model, Qwen2.5-VL-72B. As illustrated in Fig. 3, the synthesis pipeline is task-aware: each sample is first routed to an affordance branch or a trajectory branch according to its annotation type, and the teacher is prompted to produce a structured reasoning trace that remains consistent with the original grounded supervision. In total, we generate 13K CoT annotations, forming VLA-CoT-13K as high-quality intermediate supervision that bridges perception and action.

Affordance tasks synthesis. For samples annotated with affordance regions, we prompt the teacher model to generate a reasoning trace that progressively aligns the language instruction with the most likely interaction region in the

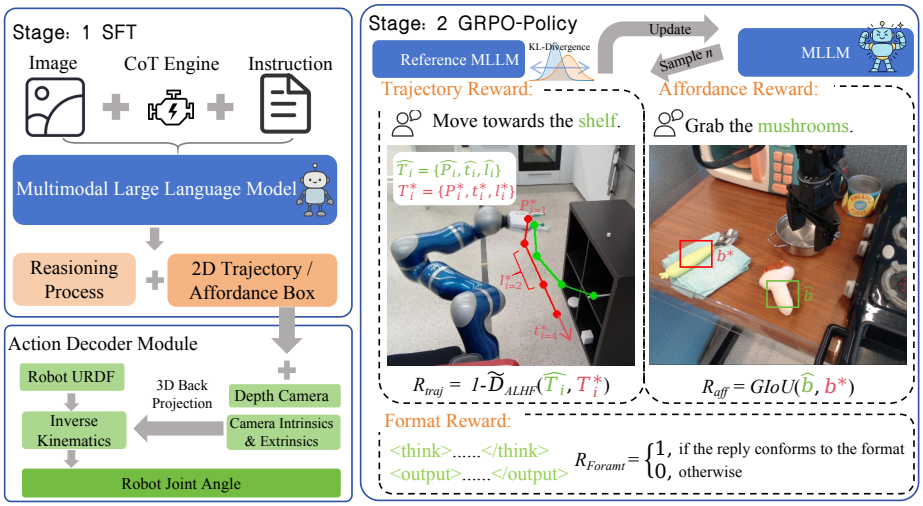

Fig. 2: Overall architecture of VLA-R1. Training has two stages: **Stage 1** uses SFT with CoT supervision to learn reasoning over images and instructions; **Stage 2** refines reasoning and actions via RL with verifiable rewards (GRPO). **During inference**, a control stack converts outputs into joint-level robot commands.

image. The process begins with comprehensive task parsing, where a complex instruction is decomposed into finer-grained atomic intentions, making explicit what interaction should be performed at the current step and what state can be regarded as successful. The teacher model then performs semantic scene understanding by identifying task-relevant entities, including the target object, other objects in the scene, and potential contact surfaces. Based on this grounded understanding, the teacher model carries out affordance region localization by highlighting the contact region most likely to support the intended manipulation, while aligning the rationale with the visual evidence. Finally, it produces a detailed feasibility check to rule out implausible candidate regions. At this stage, the reasoning explicitly considers the current object state and environmental context, such as reachability, occlusion, and whether the selected region can truly support the intended interaction. As a result, the generated CoT for this branch not only explains why a particular affordance region is correct under the given instruction, but also remains visually grounded in the original affordance supervision.

Trajectory tasks synthesis. For samples annotated with end-effector trajectories, we guide the teacher model to synthesize a reasoning process that explains how an instruction is transformed into a temporally coherent motion plan. The process first performs instruction-to-motion task parsing, where a long-horizon instruction is rewritten into a sequence of sub-trajectory objectives. It then conducts target object and gripper localization by identifying the initial state of

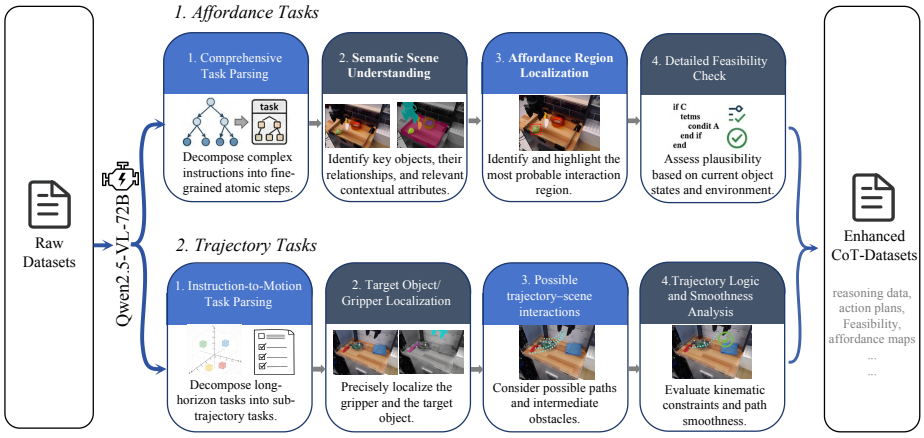

Fig. 3: Embodied CoT Data Engine. After ingesting multimodal data, the system parses tasks based on type (e.g., affordance or trajectory), performs scene understanding and localization, validates feasibility, and generates structured CoT traces for training.

the robot gripper and the positions of task-relevant target entities. Next, the teacher model reasons about possible trajectory–scene interactions, including the placement of intermediate waypoints, potential collision risks, and how the end effector should bypass obstacles while satisfying task constraints. Finally, the reasoning trace concludes with trajectory logic and smoothness analysis, which checks whether the proposed motion is kinematically feasible and smooth, avoids unrealistic discontinuities, and ensures that the path progresses coherently toward the goal. The generated CoT for this branch therefore not only specifies where the trajectory should go, but also explains why these intermediate paths are reasonable under the given scene layout.

3.3 Supervised Fine-Tuning

We perform supervised fine tuning on our synthetic high quality *VLA-CoT-13K* dataset, which presents step by step think chains paired with grounded visual evidence and action targets. Compared with naive question and answer instruction tuning, chain of thought provides intermediate supervision signals that encourage explicit decomposition, stronger visual grounding, and stable credit assignment across time. This produces policies that reason before acting, which improves sample efficiency and prepares the model for subsequent post training under verifiable rewards. In practice we supervise both the structured `<think>` segment and the final `<output>` or action segment, which regularizes reasoning style, reduces hallucination, and yields more reliable action decoding under long horizon inputs.

We initialize our foundation model with Qwen2.5-VL-3B [5]. The vision pathway is a redesigned Vision Transformer with window attention and 2D RoPE

that supports native input resolution and dynamic frame rate sampling for videos. Visual tokens are softly compressed by an MLP merger before being fed into the language decoder. The text side adopts the Qwen2.5 tokenizer with a large vocabulary and the standard Qwen2.5 decoder stack. On top of the multimodal decoder we attach an action decoder module that we implement to map hidden states to control outputs for downstream tasks. This initialization provides a strong balance of accuracy and efficiency for long temporal contexts.

3.4 Reinforcement Learning

After SFT, we further optimize VLA-R1 through RL, as shown in Fig. 2. We adopt the GRPO algorithm, recently proposed by DeepSeek [12, 34] as a scalable variant of RLHF. We extend this approach to multimodal action reasoning, allowing the model to benefit from structured verifiable rewards while maintaining training stability. For input q, GRPO samples $\{o_1, \ldots, o_n\}$ from π_{old}, scores each with a reward function to get r_g. Normalize via intra-group mean $\bar{r}$ and std σ_r: $\hat{A}_g = (r_g - \bar{r})/\sigma_r$. For process supervision, step-wise rewards are normalized similarly, with token-wise advantages accumulated and shared across outputs. For the k-th token of the g-th output, the new/old policy probability ratio is:

$$r_{g,k}(\theta) = \frac{\pi_\theta(o_{g,k} \mid q, o_{g,<k})}{\pi_{\text{old}}(o_{g,k} \mid q, o_{g,<k})}. \tag{1}$$

GRPO's objective:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = -\sum_{g=1}^{n} \frac{1}{|o_g|} \sum_{k=1}^{|o_g|} \Big[\min\Big(& r_{g,k}(\theta)\, \hat{A}_{g,k}, \\ & \text{clip}\big(r_{g,k}(\theta),\, 1-\varepsilon,\, 1+\varepsilon\big)\, \hat{A}_{g,k} \Big) \;-\; \beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \Big]. \end{aligned} \tag{2}$$

where clip$(\cdot)$ bounds the ratio to $[1-\varepsilon, 1+\varepsilon]$, and the last term is a KL penalty to avoid excessive policy drift.

Fréchet Trajectory Reward. The primary reward measures alignment using Angle-Length Augmented Fréchet distance (ALAF). Unlike pointwise Euclidean losses, ALAF respects the temporal ordering of the curves and augments it with local geometry. We represent each trajectory as a sequence of triples $T = \{p_i, t_i, \ell_i\}$, where p_i is the 2D waypoint (normalized image coordinates), t_i is the *unit* motion direction at p_i (forward/backward difference at endpoints and a normalized blend of adjacent segment directions for interior vertices), and ℓ_i is the local segment length (distance to the next waypoint; for the last vertex, to the previous one). ALAF combines the positional discrete Fréchet term with an *angle* penalty between unit tangents and a *scale* penalty based on the log ratio of neighboring segment lengths, weighted by λ_θ and λ_r; see Eq. (3).

$$D_{\mathrm{ALAF}}(\hat{T}, T^*) = \min_{\Phi} \max_{(i,j)\in\Phi} \Big[\underbrace{\|\hat{p}_i - p_j^*\|_2}_{\text{position}} + \lambda_\theta \underbrace{\arccos\Big(\frac{\hat{t}_i t_j^*}{\|\hat{t}_i\|\,\|t_j^*\|}\Big)}_{\text{angle}} + \lambda_r \underbrace{\big|\log(\hat{\ell}_i/\ell_j^*)\big|}_{\text{length ratio}} \Big], \tag{3}$$

where Φ denotes all order-preserving couplings between the sequences. $\hat{T} = \{\hat{p}_i, \hat{t}_i, \hat{\ell}_i\}$ denotes the ground-truth trajectory. $T^* = \{p_i^*, t_i^*, \ell_i^*\}$denotes the predicted trajectory. The trajectory reward is defined as

$$R_{\mathrm{traj}} = 1 - \tilde{D}_{\mathrm{ALAF}}(\hat{T}, T^*). \tag{4}$$

Here, $\tilde{D}_{\mathrm{ALAF}}$ denotes the ALAF distance normalized to $[0, 1]$; larger R_{traj} indicates better alignment.

GIoU Affordance Reward. For spatial grounding, we introduce a GIoU [33] reward between predicted and ground-truth bounding boxes. While IoU only considers the overlapping region, GIoU additionally accounts for the smallest enclosing box, penalizing misaligned predictions even when boxes do not overlap. This improves spatial robustness, especially in cluttered environments where partial overlaps are common:

$$R_{\mathrm{GIoU}} = \mathrm{GIoU}(\hat{b}, b^*). \tag{5}$$

Format Reward. Finally, we enforce structural correctness with a format reward. The model must output responses that follow the required structure (`<think>...</think>` reasoning segment followed by a `<output>...</output>` action segment). The format reward is binary:

$$R_{\mathrm{format}} = \begin{cases} 1 & \text{if the output adheres to format,} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

This encourages interpretable reasoning traces and prevents degenerate outputs during post-training.

4 Experiment

To rigorously evaluate the effectiveness and generalization capacity of the proposed approach, we conduct experiments across 4 settings: in-domain datasets, out-of-domain datasets, simulation environments, and real-robot platforms. We compare with representative baselines and ablate each component to show its impact.

4.1 Dataset and Metrics

In-Domain Datasets. We evaluate all models on ShareRobot [19], a curated subset of Open X-Embodiment [31] that covers 12 robots and 102 tasks. The benchmark includes 6,522 images with affordance annotations and 6,870 with trajectory annotations, all human-verified for quality. We further use the Embodied CoT engine to generate the VLA-CoT dataset on these subsets for training.

Out-of-Domain Datasets. For out-of-domain evaluation, we use a subset of UMD Part Affordance [29] with 1,200 samples from four categories (grasp, cut, pound, scoop). For trajectory prediction, we evaluate on VAIT [30], selecting 500 samples and manually correcting deviated trajectories for reliable assessment.

Metrics. For affordance perception, we use IoU (%) to measure overlap between predicted and ground-truth regions. For trajectory prediction, following prior work [19, 36], trajectories are represented as ordered 2D waypoints normalized to [0, 1000). We evaluate using three metrics: Discrete Fréchet Distance (DFD) for shape and alignment, Hausdorff Distance (HD) for maximum deviation, and Root Mean Square Error (RMSE) for average error, which providing a comprehensive assessment of accuracy and consistency.

In both real-world and simulated evaluations, we report Success Rate (SR) as the task-level metric. For affordance tasks, success requires correct object localization and grasp when an object is present; if no object exists, success is defined as outputting no box. For trajectory tasks, success requires ending in the goal region with the object delivered.

4.2 Experiment on Benchmark

Implementation Details. To ensure a fair comparison, we curate a contemporary suite of baselines. Specifically, we evaluate Gemma-3-12B-it, Gemma-3-27B-it [37], Phi-4-multimodal-Instruct [1], and the Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-32B-Instruct [6]. All open-source models are assessed under few-shot prompting to furnish a minimal perception prior. To validate the effectiveness of our training framework, we further include supervised fine-tuning baselines—InternVL2-2B [8] , LLaVA-1.6-7B [26], RoboBrain-7B [19] , and NORA-3B [17]—as well as an RL post-trained model, ManipLVM-R1-3B [36] and Embodied-R1-3B [45].

Experiment Results. As shown in the Table 2, open-source multimodal instruction-following models perform poorly on the in-domain dataset: despite large parameter counts, IoU remains below 10, and trajectory errors (DFD, HD, RMSE) are uniformly high. This indicates that generic models alone are inadequate for the precision demands of embodied tasks. Supervised fine-tuning (SFT) yields clear gains—e.g., RoboBrain-7B and NORA-3B attain higher IoU and lower trajectory errors than open-source baselines—yet their IoU typically remains in the 5–25 range. By contrast, VLA-R1-3B achieves the best results across all metrics: IoU = 36.51, DFD = 106.2, HD = 97.9, and RMSE = 71.12. Relative to the strong baseline ManipLVM-R1, IoU improves by 17.78%, and the

Table 2: In-domain and out-of-domain performance comparison. Avg is computed as the mean of DFD, HD, and RMSE.

Method	In-Domain					Out-of-Domain				
	IoU ↑	DFD ↓	HD ↓	RMSE ↓	Avg ↓	IoU ↑	DFD ↓	HD ↓	RMSE ↓	Avg ↓
Open-source Models										
Phi-4-multimodal-Instruct	0.58	243.92	224.73	189.27	219.31	2.17	240.18	235.44	202.69	226.10
Gemma-3-12b-it	1.18	206.72	190.64	154.96	184.11	4.65	204.94	209.88	175.42	196.75
Gemma-3-27b-it	1.32	257.42	230.29	184.47	224.06	8.20	232.86	268.03	209.25	236.71
Qwen2.5-VL-3B-Instruct	6.15	208.02	179.12	144.14	177.09	23.96	211.80	205.00	140.49	185.76
Qwen2.5-VL-32B-Instruct	7.40	125.54	113.00	85.05	107.86	25.14	182.73	176.51	133.17	164.14
Supervised Fine-Tuning										
LLaVA-1.6-7B	3.98	184.40	178.00	133.28	165.23	5.90	170.88	167.10	160.79	166.26
InternVL2-2B	6.74	250.20	239.34	194.74	228.09	15.25	165.98	167.84	145.64	159.82
RoboBrain-7B	11.79	156.10	136.52	106.71	133.11	22.00	220.94	214.14	173.02	202.70
NORA-3B	23.48	139.65	126.76	92.97	119.79	22.44	154.81	129.84	95.65	126.77
Supervised Fine-Tuning + Reinforcement Learning										
ManipLVM-R1-3B	31.00	134.18	111.14	87.28	110.87	28.00	146.82	140.52	108.64	131.99
Embodied-R1-3B	–	132.37	113.65	86.22	110.75	–	146.24	132.80	99.36	126.13
VLA-R1-3B	**36.51**	**106.20**	**97.90**	**71.12**	**91.74**	**33.96**	**114.30**	**98.43**	**68.97**	**93.90**

overall trajectory error is reduced by 17.25%, attesting to the effectiveness of our training paradigm.From the OOD results, despite substantial distribution shifts, VLA-R1-3B remains superior on trajectory prediction: IoU increases to 33.96, while DFD, HD, and RMSE decrease to 114.3, 98.43, and 68.97, respectively—surpassing the baseline, ManipLVM-R1-3B, and demonstrating strong cross-domain generalization and robustness.

4.3 Experiment on the Real World

To comprehensively assess real-world performance, we design four canonical scenarios on a tabletop platform. We instantiate: **S1: Bowl picking**, containing bowls of multiple colors placed in diverse locations; the model must grasp the user-specified color and, for trajectory tasks, place it precisely into a designated frame/basket of a given color. **S2: Fruit picking**, featuring repeated instances of the same fruit; the model must disambiguate and grasp the specified item and, for trajectory tasks, place it into the basket or onto the plate indicated by the instruction. **S3: Kitchen scenario**, comprising an open microwave, plates, and food props, where the model must contend with visual occlusion from the door and the spatial constraints of the cavity. **S4: Mixed scenario**, in which bowls, produce, baskets, and plates co-occur, requiring grasp-and-place under multi-category, multi-attribute distractors. Each scenario is evaluated over ten independent trials; we randomize initial object placements and poses and shuffle scenario order to mitigate potential ordering effects. Since Pi0/Pi0.5, OpenVLA, and RT-2-X cannot be directly deployed on an arbitrary robot embodiment without post-training adaptation, we additionally fine-tune these baselines on our collected trajectories. Specifically, we collect 160 real-world task trajectories (40 per scenario) and 80 simulated trajectories (40 per robot embodiment) for training, and fine-tune each model until the training loss converges.

Fig. 4: Evaluate in 4 real - world scenarios and 2 simulation environments

Experiment Results. As shown in Table 3, VLA-R1 achieves an average SR of 62.5% for affordance perception and 75% for trajectory prediction across the four scenarios. In comparison, OpenVLA and RT-2-X lag behind with average SRs below 40% on both tasks. More recent baselines show stronger performance: Pi0 reaches 45% and 50%, while Pi0.5 attains 62.5% and 70%, respectively. Nevertheless, VLA-R1 further improves trajectory prediction to 75% while matching the best affordance accuracy.

Across scenarios, failures are mainly caused by distractors such as color similarity, visually identical objects, and heavy clutter, especially in S2 and S4 where fine-grained discrimination and spatial reasoning are required. Despite these challenges, VLA-R1 remains stable across scenarios and demonstrates stronger spatial consistency, often producing affordance regions or trajectories close to the target even under severe clutter. These results indicate improved grounding and partial self-correction enabled by explicit reasoning supervision.

4.4 Experiment on Simulation

Implementation Details. To evaluate cross-robot performance and improve reproducibility, we conduct additional experiments in the RoboTwin simulator. A randomized tabletop clutter generator varies object categories, colors, poses, and table appearance across trials. Two robotic embodiments, Piper and UR5, are evaluated to test the cross-robot generalization ability of VLA-R1, each over ten independent trials with randomized initialization.

Experiment Results. Although trained only on real-world data, VLA-R1 performs well in simulation. On affordance perception, it achieves 60% SR on Piper and 50% on UR5; for trajectory execution, it reaches 80% and 60%, respectively, as shown in Table 3.

Earlier VLA baselines such as OpenVLA and RT-2-X perform substantially worse, with affordance and trajectory SRs generally below 30%. Stronger recent models improve performance: Pi0 achieves 40% and 30% on affordance and 50% and 40% on trajectory for Piper and UR5, while Pi0.5 reaches 50% and 50% on affordance and 70% and 60% on trajectory. NORA attains moderate affordance

Table 3: Performance comparison in real-world and simulation experiments.

Model	Task	Real World					Simulation		
		S1	S2	S3	S4	Avg	Piper	UR5	Avg
OpenVLA [21]	affordance	30%	30%	20%	20%	25%	20%	10%	15.0%
	trajectory	20%	30%	30%	40%	30.0%	20%	0%	10.0%
RT-2-X [38]	affordance	40%	40%	30%	20%	32.5%	30%	20%	25.0%
	trajectory	30%	40%	40%	50%	40.0%	30%	10%	20.0%
NORA [17]	affordance	40%	30%	30%	40%	35%	50%	30%	40%
	trajectory	40%	50%	30%	70%	47.5%	30%	10%	20%
Pi0 [18]	affordance	50%	40%	50%	40%	45%	40%	30%	35.0%
	trajectory	40%	50%	50%	60%	50.0%	50%	40%	45.0%
Pi0.5 [18]	affordance	70%	60%	60%	60%	**62.5%**	50%	50%	50.0%
	trajectory	60%	70%	80%	70%	70.0%	70%	60%	65.0%
VLA-R1	affordance	80%	60%	70%	60%	**62.5%**	60%	50%	**55%**
	trajectory	60%	80%	80%	80%	**75%**	80%	60%	**70%**

Table 4: Real-world performance comparison between Embodied-R1 and VLA-R1 over five scenarios, with 30 independent trials per scenario.

Model	Task	S1	S2	S3	S4	S5	Avg
Embodied-R1	Aff.	73.3%	60.0%	66.7%	63.3%	53.3%	63.3%
	Traj.	63.3%	73.3%	76.7%	73.3%	53.3%	68.0%
VLA-R1	Aff.	76.7%	63.3%	73.3%	66.7%	63.3%	**68.7%**
	Traj.	66.7%	76.7%	80.0%	76.7%	56.7%	**71.3%**

accuracy (50% on Piper and 30% on UR5) but remains weak on trajectory prediction. Overall, VLA-R1 achieves the highest trajectory success rate across both robots while maintaining competitive affordance performance.

Following the reviewers' suggestion, we further extend the real-world evaluation by adding S5 and increasing each scenario to 30 trials. S5 is an ambiguity-focused scene consisting of multiple colored blocks and distractor baskets, where the robot must identify the target object according to color and spatial relations rather than relying on object category alone. Since the additional S5 evaluation mainly targets ambiguity reasoning, we report the comparison between Embodied-R1 and VLA-R1 in Table 4.

4.5 Ablation Study

Implementation Details. To rigorously assess the impact of Chain-of-Thought (CoT) reasoning and Reinforcement Learning (RL) on performance, we conduct an ablation study with four configurations: (1) without CoT and RL (w/o CoT

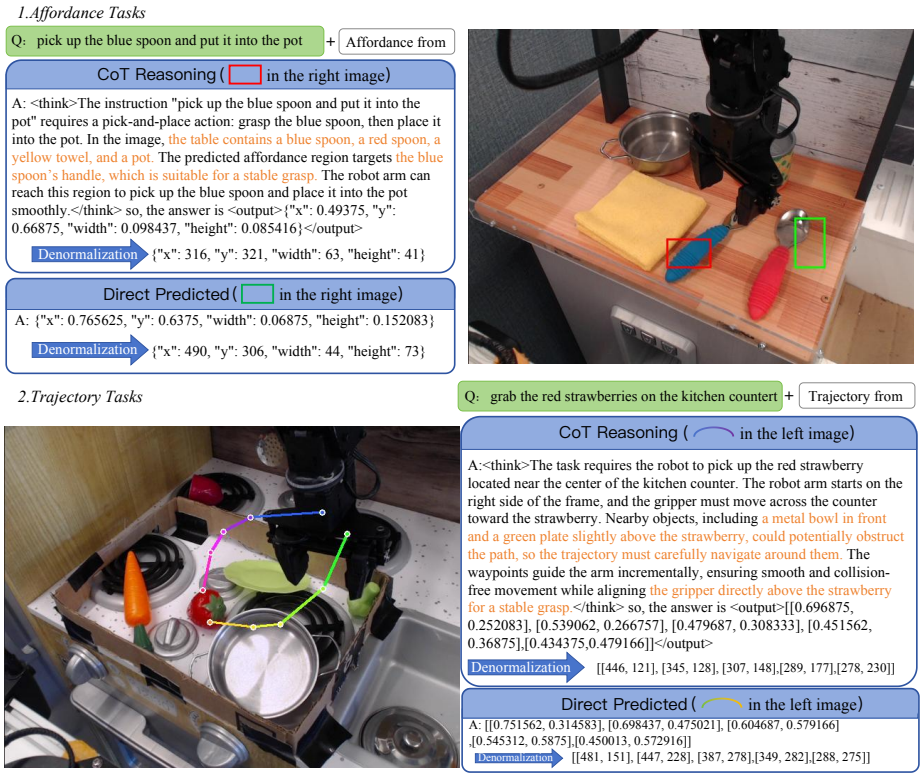

Fig. 5: Ablation Case Analysis: The figure presents a comparison between direct outputs and those generated with embodied CoT reasoning.

& RL); (2) CoT only; (3) RL only; and (4) CoT+RL. All models are trained under identical hyperparameters to ensure a fair comparison.

Experiment Results. As shown in Table 5, introducing embodied CoT reasoning alone improves affordance localization quality. IoU increases from 23.74 to 28.37, while the overall error decreases from 128.38 to 124.60. This suggests that explicit reasoning helps the model better interpret task instructions and reduce ambiguities induced by visual attributes such as color and interaction regions, which is especially important in cluttered scenes. As illustrated in Fig. 5, for affordance prediction the model can handle complex tabletop arrangements with multiple semantically similar objects (e.g., several spoons) by distinguishing the correct instance via handle color and focusing on the graspable part, rather than relying on a direct gauss. For trajectory prediction, when potential obstacles are present, the reasoning trace supports more accurate localization of both the end effector and the target object, leading to trajectories that better avoid collisions.

With reinforcement learning further applied, the RL-only variant reduces the overall error to 108.66, showing that RL can directly improve executable output

Table 5: Ablation study on the effect of Embodied CoT reasoning and RL. Avg is computed as the mean of DFD, HD, and RMSE.

Method	IoU ↑	DFD ↓	HD ↓	RMSE ↓	Avg ↓
w/o Embodied CoT & RL	23.74	149.38	135.72	100.04	128.38
Embodied CoT only	28.37	145.51	131.26	97.03	124.60
RL only	31.85	124.90	116.35	84.72	108.66
Embodied CoT + RL	**36.51**	**106.20**	**97.90**	**71.12**	**91.74**

optimization. When combined with Embodied CoT reasoning, performance improves substantially across all metrics. IoU rises to 36.51, while DFD, HD, and RMSE drop to 106.20, 97.90, and 71.12, respectively, reducing the overall error to 91.74. These results indicate that RL refines the policy using task-success signals, while Embodied CoT provides structured reasoning priors, making the predicted regions and motions better aligned with successful execution.

5 Limitation and Future Work

While VLA-R1 demonstrates strong performance across benchmarks, simulation, and real-robot settings, a key limitation is that it has not yet been developed or validated on other types of robotic platforms such as bi-manual robot arms and quadruped robot dogs. Extending VLA-R1 to these embodiments represents an important direction for future work, enabling broader applicability and testing its generalization in more diverse real-world scenarios.

6 Conclusion

In this work, we introduce **VLA-R1**, a reasoning-enhanced Vision–Language–Action model that integrates chain-of-thought supervision with reinforcement learning from verifiable rewards. By designing the **VLA-CoT-13K** dataset and incorporating an RLVR-based post-training strategy, VLA-R1 explicitly strengthens both step-by-step reasoning and execution robustness. Comprehensive experiments across in-domain, out-of-domain, simulation, and real-robot platforms demonstrate that VLA-R1 achieves SOTA performance and superior generalization. We believe this work provides a promising step toward bridging the gap between reasoning quality and action execution in embodied AI.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant No. 2021ZD0200400, the National Natural Science Foundation of China under Grant No. 62073317, and the Beijing Natural Science Foundation Haidian Original Innovation Joint Fund Project under Grant Nos. L232035 and L222154.

References

1. Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R.J., Javaheripi, M., Kauffmann, P., et al.: Phi-4 technical report. arXiv preprint arXiv:2412.08905 (2024)
2. Andrychowicz, M., Baker, B., Chociej, M., et al.: Learning dexterous in-hand manipulation. International Journal of Robotics Research **39**(1), 3–20 (2020)
3. Azzolini, A., Brandon, H., Chattopadhyay, P., et al.: Cosmos-reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558 (2025)
4. Bai, J., Bai, S., Chu, Y., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Bai, S., Chen, K., Liu, X., Wang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
7. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 (2023)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
9. Garrett, C.R., Chitnis, R., Holladay, R., Kim, B., Silver, T., Kaelbling, L.P., Lozano-Pérez, T.: Integrated task and motion planning. Annual review of control, robotics, and autonomous systems **4**(1), 265–293 (2021)
10. Geng, Y., An, B., Geng, H., et al.: Rlafford: End-to-end affordance learning for robotic manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 5880–5886 (2023)
11. Gu, S., Holly, E., Lillicrap, T., Levine, S.: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 3389–3396. IEEE (2017)
12. Guo, D., Yang, D., Zhang, H., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
13. Huang, T., Zhang, Z., Zhang, R., Zhao, Y.: Dc-scene: Data-centric learning for 3d scene understanding. arXiv preprint arXiv:2505.15232 (2025)
14. Huang, T., Zhang, Z., et al.: 3d coca: Contrastive learners are 3d captioners. arXiv preprint arXiv:2504.09518 (2025)
15. Huang, W., Wang, C., Zhang, R., Li, Y., et al.: Voxposer: Composable 3d value maps for robotic manipulation with language models. In: Proceedings of the Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 229, pp. 540–562 (2023)
16. Huang, W., Jia, B., Zhai, Z., Cao, S., et al.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
17. Hung, C.Y., Sun, Q., Hong, P., Zadeh, A., Li, C., Tan, U., Majumder, N., Poria, S., et al.: Nora: A small open-sourced generalist vision language action model for embodied tasks. arXiv preprint arXiv:2504.19854 (2025)
18. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: π0. 5: a vision-language-action model with open-world generalization, 2025. arXiv preprint arXiv:2504.16054 (2025)

19. Ji, Y., Tan, H., Shi, J., Hao, X., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. arXiv preprint arXiv:2502.21257 (2025)
20. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
21. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
22. Li, B., Zhang, Y., Guo, D., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
23. Li, X., Zhang, M., Geng, Y., et al.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18061–18070 (2024)
24. Li, X., Liu, M., Zhang, H., Yu, et al.: Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378 (2023)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (2024)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36 (2024)
27. Liu, J., Liu, M., Wang, Z., Lee, et al.: Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. arXiv preprint arXiv:2406.04339 (2024)
28. Liu, S., Wu, L., Li, B., et al.: Rdt-1b: a diffusion foundation model for bimanual manipulation. arXiv preprint arXiv:2410.07864 (2024)
29. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 1374–1381. IEEE (2015)
30. Niu, D., Sharma, Y., Biamby, G., Quenum, J., et al.: Llarva: Vision-action instruction tuning enhances robot learning. arXiv preprint arXiv:2406.11815 (2024)
31. O'Neill, A., Rehman, A., Maddukuri, A., et al.: Open X-Embodiment: Robotic learning datasets and RT-X models. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6892–6903. IEEE (2024)
32. Peng, Y., Zhang, G., Zhang, M., et al.: Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 (2025)
33. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
34. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
35. Shen, H., Liu, P., Li, J., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
36. Song, Z., Ouyang, G., Li, M., et al.: Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. arXiv preprint arXiv:2505.16517 (2025)
37. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report (2025)

38. Vuong, Q., Levine, S., Walke, H.R., Pertsch, K., Singh, A., Doshi, R., Xu, C., Luo, J., Tan, L., Shah, D., et al.: Open x-embodiment: Robotic learning datasets and rt-x models. In: Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023 (2023)
39. Wan, W., Geng, H., Liu, et al.: Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3891–3902 (2023)
40. Wang, B., Meng, X., Wang, X., Zhu, Z., Ye, A., Wang, Y., Yang, Z., Ni, C., Huang, G., Wang, X.: Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling. arXiv preprint arXiv:2507.05198 (2025)
41. Wang, P., Bai, S., Tan, S., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
42. Wang, Q., Zhang, H., Deng, C., You, Y., et al.: Sparsedff: Sparse-view feature distillation for one-shot dexterous manipulation. arXiv preprint arXiv:2310.16838 (2023)
43. Yan, H., Zhong, Z., Zhu, J., He, J., Yuan, W., Song, W., Gong, X., Cai, Y., Zhao, G., Yan, X., et al.: S-vam: Shortcut video-action model by self-distilling geometric and semantic foresight. arXiv preprint arXiv:2603.16195 (2026)
44. Ye, A., Song, Y., Su, J., Zhang, D.: Non-invasive spatial registration using customized dental bracket and improved genetic algorithms. In: 2024 IEEE 20th International Conference on Automation Science and Engineering (CASE). pp. 3250–3255. IEEE (2024)
45. Yuan, Y., Cui, H., Huang, Y., Chen, Y., Ni, F., Dong, Z., Li, P., Zheng, Y., Tang, H., Hao, J.: Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. arXiv preprint arXiv:2508.13998 (2025)
46. Zhang, W., Wang, M., Liu, G., et al.: Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. arXiv preprint arXiv:2503.21696 (2025)