

ManipLVM-R1: Reinforcement Learning for Reasoning in Embodied Manipulation with Large Vision-Language Models

Zirui Song^{1*}, Guangxian Ouyang^{1*}, Mingzhe Li², Yuheng Ji³, Chenxi Wang¹, Zixiang Xu¹, Zeyu Zhang⁴, Xiaoqing Zhang⁵, Qian Jiang¹, Fengxian Ji¹, Zhenhao Chen¹, Zhongzhi Li³, Xiuying Chen^{1†}

¹ Mohamed bin Zayed University of Artificial Intelligence

² ByteDance

³ Institute of Automation, Chinese Academy of Sciences

⁴ The Australia National University

⁵ Renmin University of China

Abstract

Large Vision-Language Models (LVLMs) have recently advanced robotic manipulation by leveraging vision for scene perception and language for instruction following. However, existing methods rely heavily on costly human-annotated training datasets, which limits their generalization and causes them to struggle in out-of-domain (OOD) scenarios, reducing real-world adaptability. To address these challenges, we propose ManipLVM-R1, a novel reinforcement learning framework that replaces traditional supervision with Reinforcement Learning using Verifiable Rewards (RLVR). By directly optimizing for task-aligned outcomes, our method enhances generalization and physical reasoning while removing the dependence on costly annotations. Specifically, we design two rule-based reward functions targeting key robotic manipulation subtasks: an Affordance Perception Reward to enhance localization of interaction regions, and a Trajectory Match Reward to ensure the physical plausibility of action paths. These rewards provide immediate feedback and impose spatial-logical constraints, encouraging the model to go beyond shallow pattern matching and instead learn deeper, more systematic reasoning about physical interactions. Experimental results show that ManipLVM-R1 achieves substantial performance gains across multiple manipulation tasks, using only 50% of the training data while achieving strong generalization to OOD scenarios. We further analyze the benefits of our reward design and its impact on task success and efficiency.

Introduction

Large Vision-Language Models (LVLMs) are driving progress toward artificial general intelligence (AGI) by combining self-supervised learning with large-scale multimodal data (Song et al. 2025; Xie et al. 2025; Wang et al. 2025; Ji et al. 2025a; Guo et al. 2024; Zhou et al. 2024). They exhibit strong visual perception and language understanding, achieving notable results in visual question answering (Cai

et al. 2024; Chen et al. 2024), image captioning (Li et al. 2022, 2023a), and mobile assistants (Song et al. 2024a; Li et al. 2024c).

Despite recent success, applying LVLMs to robotic manipulation remains limited due to their lack of fine-grained control and grounded physical reasoning. Most prior work focuses on high-level planning (Belkhale et al. 2024; Huang et al. 2023b; Song et al. 2024b), action sequencing (Brohan et al. 2023b,a), or visual reasoning (Liu, Bahety, and Song 2023; Tan et al. 2025; Zhang et al. 2025; Zhao et al. 2025; Ji et al. 2025b), with limited attention to low-level decision-making and physical dynamics required in manipulation.

RoboBrain (Ji et al. 2025c) fine-tunes LVLMs with reasoning-augmented supervision by combining expert demonstrations and language model-generated thoughts. While this improves instruction alignment, supervised fine-tuning (SFT) demands extensive annotations and suffers from limited generalization to unseen scenarios, making it less scalable for real-world deployment.

Reinforcement Learning with Verifiable Rewards (RLVR) has recently emerged as a promising alternative in domains such as mathematics and programming (Team et al. 2025b; Huang et al. 2025b,c; Chen et al. 2025; Han et al. 2024), where task outcomes can be automatically verified without human intervention. However, directly applying RLVR to robotic manipulation poses challenges: rewards are often sparse and delayed, and conventional scalar feedback lacks the spatial and logical structure necessary to guide visuomotor learning effectively.

To address these limitations, we propose ManipLVM-R1, a novel RLVR-based framework that enables LVLMs to perform robotic manipulation by optimizing structured, verifiable rewards. Specifically, we introduce task-aligned reward functions for two key subtasks: affordance perception and trajectory prediction. The affordance reward integrates format validation with IoU-based spatial matching, while the trajectory reward aggregates path similarity, endpoint precision, and movement smoothness constraints. These reward formulations provide immediate, structured feedback that is both spatially grounded and semantically aligned with phys-

*These authors contributed equally.

†Corresponding Author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

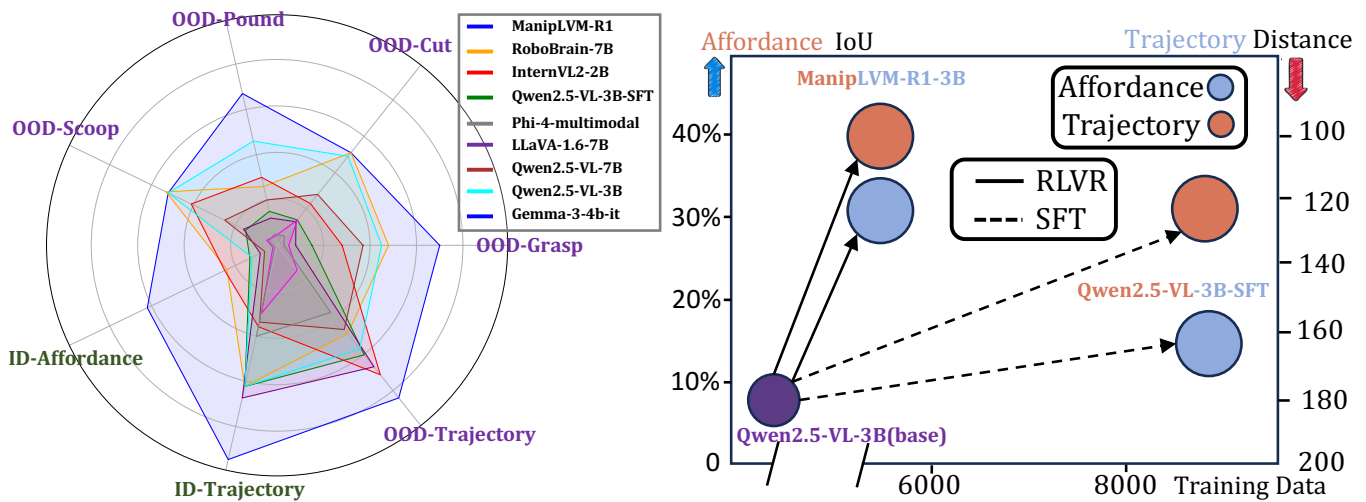


Figure 1: **Left:** Comparative evaluation of ManipLVM-R1 against baselines on **In Domain (ID)** dataset and **Out of Domain (OOD)** dataset. **Right:** Leveraging our proposed RLVR method, ManipLVM-R1 outperforms supervised fine-tuning on both affordance perception and trajectory prediction, using only 50% of the training data.

ical task success.

Experimental results show that ManipLVM-R1 outperforms strong baselines using only 50% of the training data, achieving a task success rate of 21.2% in affordance prediction and 28.1% in trajectory execution. It also yields a 144% improvement in IoU and a 12.5% gain in trajectory prediction accuracy over the strongest supervised baseline. In out-of-distribution settings, ManipLVM-R1 consistently surpasses both open-source and fine-tuned models, including Qwen2.5-VL-32B-Instruct and RoboBrain-7B (Ji et al. 2025c).

Our contributions are threefold: (1) a reinforcement learning framework for robotic manipulation that avoids reliance on costly human supervision; (2) structured, multi-objective reward functions tailored to spatial and temporal reasoning; and (3) thorough evaluations demonstrating improved sample efficiency, generalization capability, and task execution quality in both in-domain and out-of-distribution settings.

Related Work

RLVR for Large Vision Language Models

Large Vision-Language Models (LVMs) have demonstrated impressive reasoning capabilities across a variety of visual tasks (Li et al. 2024a; Wang et al. 2024a; Bai et al. 2023; Wang et al. 2024b). Their ability to align visual perception with language understanding has enabled strong performance in domains such as question answering, captioning, and instruction following.

To further enhance these capabilities, recent research has explored Reinforcement Learning with Verifiable Rewards (RLVR). Vision-R1 (Huang et al. 2025a) introduces the PTST strategy, allowing effective training on cold-start datasets while matching performance with significantly larger models. LMM-R1 (Peng et al. 2025) employs a two-stage RLVR pipeline, leveraging text-only reasoning in the first stage before transferring to vision-based domains,

achieving effective modality transfer. VLM-R1 (Shen et al. 2025) adapts the RLVR paradigm to open-vocabulary object detection by grounding language in visual entities through verifiable feedback. These approaches show that RLVR can effectively activate latent reasoning capabilities within pre-trained models, even without extensive human supervision.

However, none of these directly address the domain of robotic manipulation, which requires not just perception and reasoning but also physical execution and control. To bridge this gap, we propose ManipLVM-R1, an RLVR-based framework tailored to manipulation tasks, integrating both high-level reasoning and low-level action grounding.

Robot Manipulation

Traditional robot manipulation methods often relied on low-level state-based reinforcement learning (Geng et al. 2023), which struggled to generalize to complex, high-dimensional visual inputs. The recent integration of LLMs and vision-language models into manipulation pipelines has led to a new wave of vision-based manipulation approaches (Brohan et al. 2022, 2023a; Wang et al. 2023; Li et al. 2024b; Liu et al. 2024c), enabling more abstract task planning and generalization.

Some methods aim to ground language into spatial representations. VoxPoser (Huang et al. 2023a) synthesizes 3D volumetric value maps from textual instructions to guide object placement. Others focus on fine-tuning VLMs for robotic tasks. RoboFlamingo (Li et al. 2023b) and ManipLLM (Li et al. 2024b) supervise models with CoT data to learn visual reasoning for manipulation. RoboMamba (Liu et al. 2024c) and OpenVLA (Kim et al. 2024) further extend this by curating high-quality datasets for language-guided visual planning.

Embodied Reasoner (Zhang et al. 2025), Cosmos-Reason1 (Azzolini et al. 2025), and RoboBrain (Ji et al. 2025c) push the boundary by improving interpretability and

reasoning alignment through supervised CoT fine-tuning.

Despite these advances, most approaches still depend heavily on large-scale annotations and handcrafted supervision strategies. In contrast, our ManipLVM-R1 framework eliminates the need for extensive CoT labels, instead learning from structured, verifiable rewards. This enables robust reasoning and generalization under low supervision conditions, setting the stage for more scalable and autonomous robot learning.

Method

In this section, we first present the preliminaries of RLVR. Subsequently, we introduce the training framework for ManipLVM-R1 and elucidate the rationale behind its design. A visual overview of our architecture is shown in Figure 2.

Preliminary of Reinforcement Learning with Verifiable Rewards

Reinforcement Learning with Verifiable Rewards (RLVR) (Guo et al. 2025; Team et al. 2025b) is a paradigm for training language or vision-language models on tasks where correctness can be automatically verified—such as in mathematics, programming, and increasingly, embodied AI. Unlike reinforcement learning from human feedback (RLHF) (Liu et al. 2024a; Zang et al. 2025), RLVR removes the need for subjective preference labels by relying on task-specific verification functions to determine reward signals.

Formally, given an instruction q , a policy model π_θ generates a candidate response o , which is then evaluated by a verifiable reward function $R(q, o)$. The model is trained to maximize the following RLVR objective:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_\theta(q)} [R(q, o) - \beta \text{KL}(\pi_\theta(o|q) \parallel \pi_{\text{ref}}(o|q))] \quad (1)$$

Here, π_{ref} is the reference model prior to RLVR optimization, and β is a hyperparameter that regulates the divergence from this reference. The reward function R provides binary or continuous signals based on whether the model’s output satisfies task-specific success criteria. For binary tasks, it typically takes the form:

$$R(q, o) = \begin{cases} 1, & \text{if } o = \text{ground truth,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In practice, more expressive reward functions can incorporate task structure and partial correctness signals—crucial for domains like robotic manipulation where outcomes are spatially and temporally grounded. Our ManipLVM-R1 leverages this flexibility by designing reward functions tailored to affordance perception and trajectory alignment, thereby guiding the model toward physically plausible and task-effective behaviors.

ManipLVM-R1

While RLVR has shown promise in domains like mathematics and programming, its application to robotic manipulation remains limited. This is primarily due to two challenges:

(1) sparse and delayed reward signals in manipulation tasks, which hinder effective learning and optimization, and (2) the lack of spatial-logical constraints in reward design, causing models to rely on superficial pattern matching rather than learning grounded physical reasoning. To address these issues, we propose ManipLVM-R1, an RLVR-based framework that incorporates structured reward design and stable policy optimization. ManipLVM-R1 decomposes the overall manipulation task into two distinct subtasks: affordance perception and trajectory prediction. It trains separate models for each subtask, enabling more focused learning and better generalization in physically grounded environments.

Policy Samples For a given input state (x, q) , where x is the visual encoding of the input image and q is the textual encoding of the instruction, ManipLVM-R1 first generates G distinct responses $\{o_1, o_2, \dots, o_G\}$ from the current policy π_θ . Different from other work, to adapt this methodology for robotic manipulation tasks involving affordance perception and trajectory prediction, ManipLVM-R1 incorporates distinct sets of reward functions: *Affordance Perception Reward* and *Trajectory Match Reward*. For brevity, we omit the response index in the following introduction.

Affordance Perception Reward For the affordance perception task, the Affordance Perception Reward R_{spatial} combines two components: a format reward R_{format} and an affordance reward R_{aff} . The format reward R_{format} ensures that the model outputs follow a structured format—reasoning enclosed in ‘ $\text{;think}_i/\text{think}_i$ ’ and the final answer enclosed in ‘ $\text{;answer}_i/\text{answer}_i$ ’, with the ‘ ;answer_i ’ field containing the coordinates of a predicted bounding box. This structure not only facilitates reliable reward extraction but also guides the model to organize its output in a way that aligns with downstream evaluation and training. Meanwhile, the affordance reward R_{aff} focuses on the core objective of affordance perception: accurately identifying where an object interaction can occur. It evaluates spatial accuracy by comparing the predicted bounding box b^* with the ground-truth box \hat{b} using the Intersection-over-Union (IoU) metric. By optimizing $R_{\text{aff}} = \text{IoU}(b^*, \hat{b})$, the model is encouraged to generate spatially precise and semantically grounded predictions that directly support successful interaction planning. Taken together, the Affordance Perception Reward is defined as:

$$R_{\text{spatial}} = R_{\text{format}} + R_{\text{aff}}. \quad (3)$$

Trajectory Match Reward For the trajectory prediction task, the model is expected to output a sequence of 2D coordinates, denoted as $\hat{T} = \{\hat{p}_1, \hat{p}_2, \dots, \hat{p}_N\}$, where each $\hat{p}_i \in \mathbb{R}^2$ represents an (x, y) position in image space, with the origin at the top-left corner. The predicted trajectory is evaluated against a ground-truth sequence $T^* = \{p_1^*, p_2^*, \dots, p_M^*\}$ using three components: a format reward R_{format} , a path similarity reward R_{path} , and an endpoint distance reward R_{end} .

The format reward R_{format} verifies structural correctness, including proper use of ‘ ;think_i ’ and ‘ ;answer_i ’ tags. Unlike in affordance perception, it enforces a trajectory-specific constraint: the predicted numbers must fall within a valid range (e.g., 3–10), ensuring both conciseness and interpretability.

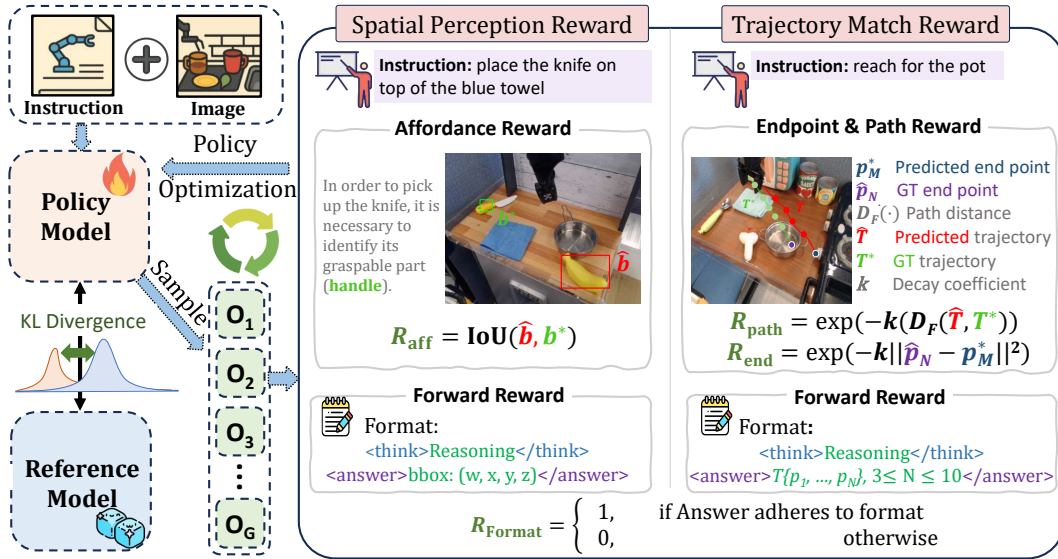


Figure 2: Overview of the proposed ManipLVM-R1 framework. Given an image and an instruction, the policy model generates multiple responses, which are optimized using structured rewards: the Affordance Perception Reward for affordance detection, and the Trajectory Match Reward based on multiple distance metrics, path similarity, and endpoint accuracy.

To encourage the model to generate physically plausible and goal-aligned trajectories, we design the path similarity reward R_{path} . The motivation behind this design is that many supervised methods, such as behavioral cloning (Nazeri and Bohlouli 2021), focus primarily on endpoint accuracy while overlooking the quality and feasibility of the trajectory itself. In robotic manipulation tasks, the shape, continuity, and alignment of the trajectory with respect to the ground truth are critical for ensuring robust and safe control strategies. To capture these aspects, we incorporate multiple geometric similarity metrics that evaluate how well the predicted trajectory matches the ground truth from different perspectives. R_{path} is a weighted aggregation of rewards derived from three distinct metrics: Discrete Fréchet Distance (DFD), Hausdorff Distance (HD), and Root Mean Square Error (RMSE). The reward for each metric (R_{DFD} , R_{HD} , R_{RMSE}) is obtained by normalizing the respective distance value to a score between 0 and 1. R_{DFD} quantifies overall shape similarity and temporal alignment between trajectories; R_{HD} captures the maximum pointwise deviation; R_{RMSE} measures the average pointwise discrepancy. The final R_{path} is as:

$$R_{\text{path}} = R_{\text{DFD}} + R_{\text{HD}} + R_{\text{RMSE}}. \quad (4)$$

The endpoint distance reward R_{end} measures the proximity between the final predicted point \hat{p}_N and the ground-truth endpoint p_M using Euclidean distance: $R_{\text{end}} = \exp(-k \|\hat{p}_N - p_M^*\|^2)$. The overall trajectory match reward is defined as the sum of the three components:

$$R_{\text{trajectory}} = R_{\text{format}} + R_{\text{path}} + R_{\text{end}}. \quad (5)$$

Policy Update Inspired by Group Relative Policy Optimization (GRPO) (Guo et al. 2025), we select multiple responses from the current policy as candidate actions. Depending on the task type, either R_{spatial} or $R_{\text{trajectory}}$ assigns

a scalar reward r_i to each response, resulting in a reward set $\{r_1, r_2, \dots, r_G\}$. To assess the quality of each response relative to others, we normalize the rewards by computing the mean and standard deviation:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \dots, r_G\})}{\text{std}(\{r_1, \dots, r_G\})}, \quad (6)$$

where A_i denotes the advantage of the i -th response. These advantages are then used to update the policy, increasing the likelihood of high-quality responses while suppressing lower-quality ones. To ensure stability during training, the update is constrained by minimizing the KL divergence between the updated policy and a reference model: $\text{KL}[\pi_\theta(o|q) \parallel \pi_{\text{ref}}(o|q)]$.

Experiment

Baselines

To establish strong performance benchmarks, we include a diverse suite of baseline models from both few-shot and supervised fine-tuning paradigms. Specifically, we evaluate three of the most recent, high-performing open-source vision-language model series: Gemma-3-4B-it, Gemma-3-12B-it, Gemma-3-27B-it (Team et al. 2025a), Phi-4-Multimodal-Instruct (Abouelenin et al. 2025), and the Qwen2.5-VL family (3B/7B/32B) (Team 2024). These models represent state-of-the-art instruction-tuned LVLMs with general visuomotor reasoning capabilities. To test their zero-shot and few-shot reasoning potential, we implemented structured few-shot prompting and ensured visual instruction coverage in the input demonstrations.

To validate the effectiveness of our reinforcement learning framework with verifiable rewards, we also compare against a range of fully supervised models. These in-

Method	In-Domain						
	Acc. \uparrow	IoU \uparrow	Acc. \uparrow	DFD \downarrow	HD \downarrow	RMSE \downarrow	Avg \downarrow
Open-source Models							
Phi-4-multimodal-Instruct	0	0.58	0.06	243.92	224.73	189.27	228.21
Gemma-3-4b-it	0	0.91	0.05	266.62	243.19	210.48	240.10
Gemma-3-27b-it	0	1.32	0.07	257.42	230.29	184.47	224.00
Gemma-3-12b-it	0	1.18	0.09	206.72	190.64	154.96	184.10
Qwen2.5-VL-3B-Instruct	0	6.15	0.12	202.86	179.12	144.14	175.37
Qwen2.5-VL-7B-Instruct	0	2.98	0.12	262.80	243.03	190.81	232.21
Qwen2.5-VL-32B-Instruct	0	7.40	<u>0.27</u>	125.54	113.00	85.05	107.86
LLARVA	0	3.95	0.16	190.12	175.64	136.78	167.51
VILA1.5-13B	0	5.15	0.18	174.87	160.88	141.56	159.10
Supervised Fine-Tuning							
LLaVA-1.6-7B	0	3.98	0.15	184.40	178.00	133.28	165.23
InternVL2-2B	0	6.74	0.11	250.20	239.34	194.74	228.09
Qwen2.5-VL-3B-Instruct	<u>0.03</u>	<u>12.69</u>	0.20	147.38	138.90	94.03	126.77
RoboBrain-7B	0.02	11.79	0.23	156.10	136.52	106.71	133.11
Our Proposed Model							
ManipLVM-R1-3B	0.21	31.0	0.28	<u>134.18</u>	<u>111.14</u>	<u>87.28</u>	<u>110.87</u>

Table 1: Comparison of model performance on the In-Domain dataset. **Bolded** values indicate best performance, and underlined values indicate second best. Task-specific metrics are color-coded: **red** for Affordance Perception and **green** for Trajectory Prediction.

clude InternVL2-2B (Chen et al. 2024), Qwen2.5-VL-3B-Instruct (Team 2024), LLaVA-1.6-7B (Liu et al. 2024b), and RoboBrain-7B (Ji et al. 2025c). Among them, RoboBrain was explicitly designed for affordance and trajectory tasks, fine-tuned from LLaVA-1.6-7B using a large-scale synthetic chain-of-thought (CoT) dataset tailored for embodied manipulation. In contrast, our method, ManipLVM-R1, achieves competitive or superior performance using only 50% of the training data, highlighting its efficiency and generalization capabilities under limited supervision.

Dataset Setting

In-Domain Dataset All baseline and proposed models are trained on the ShareRobot (Ji et al. 2025c) dataset, a large-scale resource curated from Open X-Embodiment (O’Neill et al. 2024). It is tailored for fine-grained robotic manipulation tasks including affordance perception and trajectory generation. The dataset includes over 1 million planning QA pairs, 6.5k images with dense affordance annotations, and 6.8k trajectory-labeled frames spanning 102 real-world scenes and 12 robotic embodiments. Rich multimodal supervision provides an ideal testbed for learning spatial and temporal reasoning. Importantly, all SFT baselines use the full training set, while ManipLVM-R1 trains on only 50% of the data, demonstrating strong sample efficiency.

Out-of-Domain Dataset To evaluate the robustness and generalization of our model, we conduct extensive out-of-domain (OOD) testing. For affordance perception, we select four semantic categories (*grasp*, *cut*, *pound*, *scoop*) from the UMD Part Affordance dataset (Nguyen et al. 2017). This dataset provides RGB-D images and 3D object point clouds covering 105 common tools across household and industrial contexts. We randomly sample 1200 test images from the four categories with equal distribution, ensuring diverse coverage in geometry and affordance semantics.

For trajectory prediction, we use a subset of 500 examples from the validation split of VAIT (Niu et al. 2024), derived from the Open X-Embodiment corpus. VAIT consists of image-trajectory pairs capturing a broad distribution of robot behaviors and environments. However, as it is automatically mined, it contains noisy or misaligned trajectory labels. To address this, we visualize all selected examples and apply manual verification and correction of annotations where necessary, ensuring that the OOD evaluation remains clean and reliable. This careful curation enables precise assessment of model performance under distribution shift and perceptual uncertainty.

Metric Setting

Affordance Perception: We report task success rate as the primary evaluation metric, where a prediction is considered successful if the Intersection over Union (IoU) with the ground truth exceeds 50%. This threshold-based criterion reflects whether the model can accurately localize the functional region relevant to the task. In addition, we include mean IoU as an auxiliary metric to capture the average spatial alignment quality, which provides finer-grained insight into partial matches and near misses.

Trajectory Prediction: Success is defined as the final predicted waypoint falling within a 20-pixel Euclidean distance of the ground truth endpoint, reflecting whether the action can reach the intended goal. To capture global trajectory similarity and path consistency, we also report three alignment metrics—Discrete Fréchet Distance (DFD), Hausdorff Distance (HD), and Root Mean Square Error (RMSE)—following (Chao et al. 2021; Wang et al. 2024a; Ji et al. 2025c). DFD measures overall shape resemblance and temporal coherence, HD identifies the worst-case deviation between two trajectories, and RMSE quantifies average pointwise error. Together, these metrics provide a comprehensive view of the model’s spatial and temporal accuracy.

Method	Out-of-Domain									
	UMD Affordance					VAIT Trajectory				
	Acc. \uparrow	Grasp-IoU \uparrow	Cut-IoU \uparrow	Pound-IoU \uparrow	Scoop-IoU \uparrow	Acc. \uparrow	DFD \downarrow	HD \downarrow	RMSE \downarrow	Avg \downarrow
Open-source Models										
Phi-4-multimodal-Instruct	0	1.49	2.53	2.44	2.22	0.07	240.18	235.44	202.69	226.10
Gemma-3-4b-it	0	2.29	7.04	2.24	2.26	0.05	295.48	289.41	232.21	272.31
Gemma-3-12b-it	0.02	4.14	5.25	4.36	4.66	0.16	204.94	200.89	175.42	193.75
Gemma-3-27b-it	0.07	10.29	2.08	6.27	4.05	0.07	273.86	268.63	209.52	250.67
Qwen2.5-VL-3B-Instruct	0.14	22.48	24.50	22.77	26.09	0.11	211.80	205.04	140.69	185.84
Qwen2.5-VL-7B-Instruct	0.13	18.64	14.19	10.07	12.31	0.10	228.04	222.92	170.94	207.30
Qwen2.5-VL-32B-Instruct	<u>0.17</u>	<u>24.67</u>	<u>25.84</u>	24.04	25.99	0.12	182.73	176.51	<u>133.17</u>	164.14
LLARVA	0.16	15.12	13.77	10.07	12.31	0.13	203.24	198.24	164.24	178.44
VILA1.5-13B	0.14	19.82	16.35	14.79	17.44	0.18	194.34	185.22	153.90	167.18
Supervised Fine-Tuning										
LLaVA-1.6-7B	0.02	3.89	6.34	6.05	7.31	0.15	170.88	167.10	160.79	166.25
Qwen2.5-VL-3B-Instruct	0.04	7.39	7.11	7.50	7.81	0.14	186.48	184.95	166.92	179.45
InternVL2-2B	0.10	13.93	11.52	15.06	20.48	<u>0.22</u>	<u>165.98</u>	<u>160.87</u>	145.64	<u>157.50</u>
RoboBrain-7B	0.16	23.86	25.37	12.75	<u>26.35</u>	0.15	220.94	214.14	173.02	202.70
Our Proposed Model										
ManipLVM-R1-3B	0.24	34.65	25.58	<u>23.50</u>	28.27	0.24	146.82	140.52	108.64	131.99

Table 2: Comparison of model performance on the out-of-domain (OOD) dataset. **Bolded** values indicate the best performance, and underlined values indicate the second best. Task success rate is used as the primary evaluation metric. IoU and trajectory alignment scores are reported as auxiliary metrics for affordance perception and trajectory prediction, respectively. **Red** highlights correspond to the Affordance Perception task, while **green** highlights represent the Trajectory Prediction task.

in trajectory generation, which is critical for ensuring both effectiveness and safety in downstream execution.

Quantitative Analysis

We evaluate the performance of ManipLVM-R1-3B on both affordance perception and trajectory prediction tasks under in-domain and OOD settings.

In the in-domain experiment, our model achieves an IoU of 31.0 on the affordance task, substantially outperforming all baselines, including the much larger RoboBrain-7B (11.79), despite using only 50% of the training data. This underscores the effectiveness of our structured reward design in generating spatially grounded and semantically meaningful predictions under limited supervision, validating our motivation to align learning signals directly with task utility.

For trajectory prediction, ManipLVM-R1 also demonstrates strong performance, with an average endpoint error of 110.87 which comparable to fully supervised, larger models indicating that our reward formulation captures core requirements for physical feasibility and goal-directed plausibility, as intended.

In the OOD experiment, ManipLVM-R1 continues to exhibit robust generalization. It achieves the highest Grasp-IoU (34.65) and competitive results across other UMD sub-tasks. More notably, it obtains the lowest average trajectory error (131.99) on the VAIT benchmark, outperforming all open-source and fine-tuned models, including Qwen2.5-VL-32B-Instruct and RoboBrain-7B.

These results collectively demonstrate ManipLVM-R1’s ability to generalize under distribution shifts while consistently maintaining high task accuracy across both affordance perception and trajectory prediction domains.

Qualitative Analysis

To qualitatively assess the capabilities of ManipLVM-R1, we present visual examples of its performance in affordance perception and trajectory prediction across both spatial and temporal reasoning tasks. To highlight the model’s real-world applicability, Figure 3 showcases its responses to two natural language commands: “Move the corn next to the blue toy” and “Wipe the table with yellow towel,” demonstrating its ability to bridge high-level language and low-level visuo-motor execution.

Beyond task completion, we analyze ManipLVM-R1’s internal reasoning process. As shown in Figure 3, certain cases demonstrate “aha moments” where the model exhibits emergent multi-step reasoning—without explicit chain-of-thought supervision. In one scenario, the model is instructed to insert a round object into its slot. It implicitly infers the need to match shape and color, then identifies the corresponding target location, suggesting latent planning and visual grounding. These examples highlight the potential of verifiable rewards to elicit structured reasoning in vision-language agents trained without costly manual annotations. The reasoning traces also serve as interpretable evidence of the model’s decision path.

We also present failure analysis in Figure 3, which reveals key limitations in the model’s perceptual and common-sense reasoning capabilities. One error arises from a commonsense gap: the model fails to associate a doorknob with interactivity and instead focuses on the keyhole, generating the flawed inference that “approaching the keyhole opens the door.” This points to a limitation in embodied knowledge grounding, particularly for infrequent object-function associations. Another failure stems from spatial misperception: the model misidentifies the gripper’s location relative to the stove, causing a trajectory prediction that violates the task’s spatial constraints. These cases suggest that despite

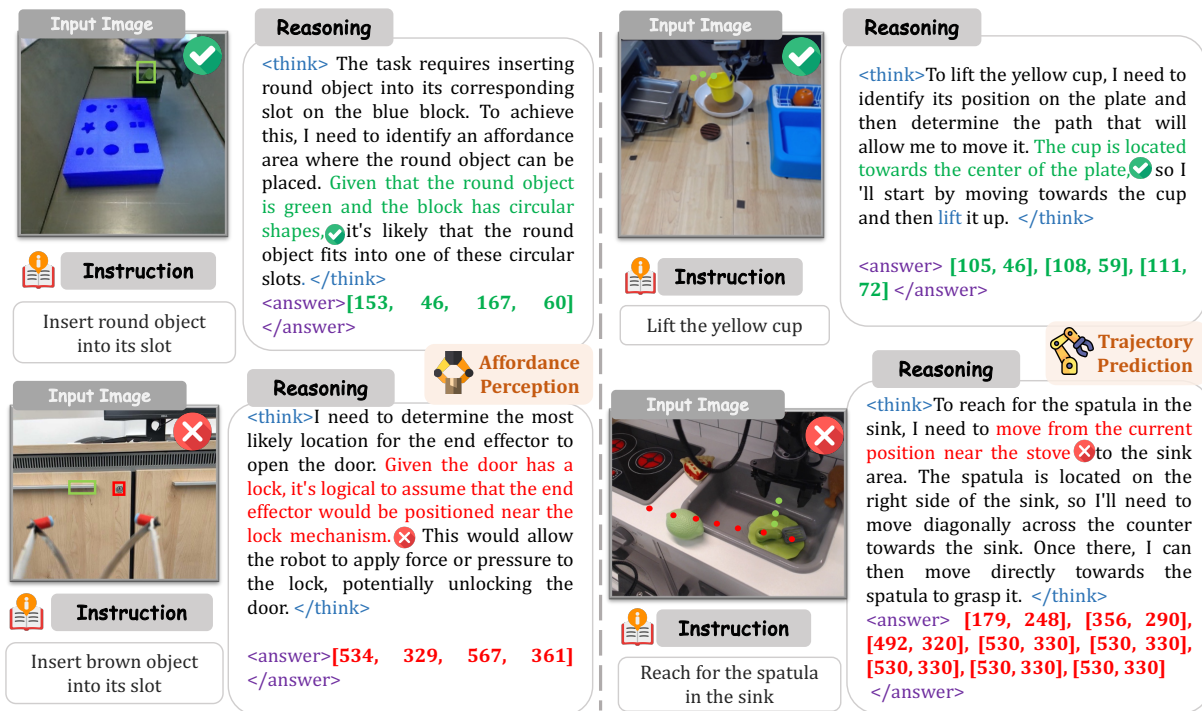


Figure 3: Case studies highlight both successful “aha moments” and failure cases in ManipLVM-R1’s predictions. Final answers appear in `<answer>` tags, with predictions or ground truth shown as green boxes and green points. Reasoning is displayed within `<think>` tags, where green text indicates correct steps and red text highlights errors.

strong performance on structured tasks, ManipLVM-R1 still lacks self-verification mechanisms and is vulnerable to compounding perceptual errors during inference.

Overall, these qualitative results underscore both the strengths and current limitations of reward-driven manipulation learning. They affirm that verifiable structured rewards can promote latent spatial reasoning and goal alignment, enabling the agent to make efficient and accurate decisions in well-defined tasks. However, these results also highlight the challenges faced when dealing with more complex, uncertain, or noisy environments. There is a clear need for enhanced introspection capabilities, such as integrating common-sense knowledge bases, uncertainty modeling, or spatial consistency checks, to ensure that the agent can robustly handle out-of-distribution or ambiguous inputs. Furthermore, future advancements could focus on incorporating adaptive feedback mechanisms to allow the model to better adjust to unexpected changes in task dynamics, further improving its performance in real-world applications.

Conclusion

We present ManipLVM-R1, a reinforcement learning framework that replaces costly human supervision with verifiable, rule-based rewards for robotic manipulation. By introducing task-aligned reward functions for affordance perception and trajectory prediction, our method enhances systematic reasoning and generalization, particularly in out-of-domain settings. Unlike prior annotation-heavy approaches,

ManipLVM-R1 uses immediate structured feedback to guide learning, enabling shorter inference paths, higher decision efficiency, and improved robustness under limited supervision. Experiments show that ManipLVM-R1 reduces supervision costs while maintaining strong performance with only 50% of the training data, illustrating a scalable direction for efficient robot learning.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. The work was supported by Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) through grant award 8481000078

References

- Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Azzolini, A.; Brandon, H.; Chattopadhyay, P.; Chen, H.; Chu, J.; Cui, Y.; Diamond, J.; Ding, Y.; Ferroni, F.; Govindaraju, R.; et al. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

- Belkhal, S.; Ding, T.; Xiao, T.; Sermanet, P.; Vuong, Q.; Tompson, J.; Chebotar, Y.; Dwibedi, D.; and Sadigh, D. 2024. Rt-h: Action Hierarchies Using Language. *arXiv preprint arXiv:2403.01823*.
- Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023a. Rt-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818*.
- Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. 2022. Rt-1: Robotics Transformer for Real-World Control at Scale. *arXiv preprint arXiv:2212.06817*.
- Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; et al. 2023b. Do as I Can, Not as I Say: Grounding Language in Robotic Affordances. In *CoRL*, 287–318.
- Cai, R.; Song, Z.; Guan, D.; Chen, Z.; Li, Y.; Luo, X.; Yi, C.; and Kot, A. 2024. Benchlm: Benchmarking cross-style visual capability of large multimodal models. In *European Conference on Computer Vision*, 340–358. Springer.
- Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9044–9053.
- Chen, X.; Wang, T.; Guo, T.; Guo, K.; Zhou, J.; Li, H.; Song, Z.; Gao, X.; and Zhang, X. 2025. Unveiling the power of language models in chemical research question answering. *Communications Chemistry*, 8(1): 4.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 24185–24198.
- Geng, Y.; An, B.; Geng, H.; Chen, Y.; Yang, Y.; and Dong, H. 2023. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5880–5886. IEEE.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Han, W.; Fang, M.; Zhang, Z.; Yin, Y.; Song, Z.; Chen, L.; Pechenizkiy, M.; and Chen, Q. 2024. MedINST: Meta Dataset of Biomedical Instructions. *arXiv preprint arXiv:2410.13458*.
- Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Hu, Y.; and Lin, S. 2025a. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Huang, W.; Wang, C.; Zhang, R.; Li, Y.; Wu, J.; and Fei-Fei, L. 2023a. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. In *CoRL*, volume 229 of *Proceedings of Machine Learning Research*, 540–562.
- Huang, W.; Xia, F.; Xiao, T.; Chan, H.; Liang, J.; Florence, P.; Zeng, A.; Tompson, J.; Mordatch, I.; Chebotar, Y.; et al. 2023b. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *CoRL*, 1769–1782. PMLR.
- Huang, Y.; Gao, C.; Wu, S.; Wang, H.; Wang, X.; Zhou, Y.; Wang, Y.; Ye, J.; Shi, J.; Zhang, Q.; et al. 2025b. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*.
- Huang, Y.; Wang, Y.; Xu, Z.; Gao, C.; Wu, S.; Ye, J.; Chen, X.; Chen, P.-Y.; and Zhang, X. 2025c. Breaking Focus: Contextual Distraction Curse in Large Language Models. *arXiv preprint arXiv:2502.01609*.
- Ji, F.; Yang, J.; Song, Z.; Wang, Y.; Cui, Z.; Li, Y.; Jiang, Q.; Fang, M.; and Chen, X. 2025a. FineState-Bench: A Comprehensive Benchmark for Fine-Grained State Control in GUI Agents. *arXiv preprint arXiv:2508.09241*.
- Ji, Y.; Liu, Y.; Zhang, Z.; Zhang, Z.; Zhao, Y.; Hao, X.; Zhou, G.; Zhang, X.; and Zheng, X. 2025b. Enhancing adversarial robustness of vision-language models through low-rank adaptation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 550–559.
- Ji, Y.; Tan, H.; Shi, J.; Hao, X.; Zhang, Y.; Zhang, H.; Wang, P.; Zhao, M.; Mu, Y.; An, P.; et al. 2025c. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257*.
- Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sankeki, P.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *EMNLP*, 7241–7259.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.
- Li, X.; Liu, M.; Zhang, H.; Yu, C.; Xu, J.; Wu, H.; Cheang, C.; Jing, Y.; Zhang, W.; Liu, H.; et al. 2023b. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.
- Li, X.; Zhang, M.; Geng, Y.; Geng, H.; Long, Y.; Shen, Y.; Zhang, R.; Liu, J.; and Dong, H. 2024b. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18061–18070.

- Li, Y.; Zhang, C.; Yang, W.; Fu, B.; Cheng, P.; Chen, X.; Chen, L.; and Wei, Y. 2024c. Appagent v2: Advanced agent for flexible mobile interactions. *arXiv preprint arXiv:2408.11824*.
- Liu, C. Y.; Zeng, L.; Liu, J.; Yan, R.; He, J.; Wang, C.; Yan, S.; Liu, Y.; and Zhou, Y. 2024a. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs. *arXiv preprint arXiv:2410.18451*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual Instruction Tuning. *NeurIPS*, 36.
- Liu, J.; Liu, M.; Wang, Z.; Lee, L.; Zhou, K.; An, P.; Yang, S.; Zhang, R.; Guo, Y.; and Zhang, S. 2024c. RoboMamba: Multimodal State Space Model for Efficient Robot Reasoning and Manipulation. *arXiv preprint arXiv:2406.04339*.
- Liu, Z.; Bahety, A.; and Song, S. 2023. REFLECT: Summarizing Robot Experiences for Failure Explanation and Correction. In *CoRL*, 3468–3484. PMLR.
- Nazeri, M. H.; and Bohlouli, M. 2021. Exploring reflective limitation of behavior cloning in autonomous vehicles. In *2021 IEEE International Conference on Data Mining (ICDM)*, 1252–1257. IEEE.
- Nguyen, A.; Kanoulas, D.; Caldwell, D. G.; and Tsagarakis, N. G. 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5908–5915. IEEE.
- Niu, D.; Sharma, Y.; Biamby, G.; Quenum, J.; Bai, Y.; Shi, B.; Darrell, T.; and Herzig, R. 2024. LLARVA: Vision-Action Instruction Tuning Enhances Robot Learning. *arXiv preprint arXiv:2406.11815*.
- O’Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandekar, A.; Jain, A.; et al. 2024. 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)*, 6892–6903. IEEE.
- Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. Lmm-r1: Empowering 3b llms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.
- Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.
- Song, Z.; Li, Y.; Fang, M.; Chen, Z.; Shi, Z.; and Huang, Y. 2024a. Mmac-copilot: Multi-modal agent collaboration operating system copilot. *arXiv e-prints*, arXiv–2404.
- Song, Z.; Ouyang, G.; Fang, M.; Na, H.; Shi, Z.; Chen, Z.; Fu, Y.; Zhang, Z.; Jiang, S.; Fang, M.; et al. 2024b. Hazards in Daily Life? Enabling Robots to Proactively Detect and Resolve Anomalies. *arXiv preprint arXiv:2411.00781*.
- Song, Z.; Yang, J.; Huang, Y.; Tonglet, J.; Zhang, Z.; Cheng, T.; Fang, M.; Gurevych, I.; and Chen, X. 2025. Geolocation with Real Human Gameplay Data: A Large-Scale Dataset and Human-Like Reasoning Framework. *arXiv preprint arXiv:2502.13759*.
- Tan, H.; Chi, C.; Chen, X.; Ji, Y.; Zhao, Z.; Hao, X.; Lyu, Y.; Cao, M.; Zhao, J.; Lyu, H.; et al. 2025. RoboOS-NeXT: A Unified Memory-based Framework for Lifelong, Scalable, and Robust Multi-Robot Collaboration. *arXiv preprint arXiv:2510.26536*.
- Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025b. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Team, Q. 2024. Qwen2.5: A Party of Foundation Models.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, Q.; Zhang, H.; Deng, C.; You, Y.; Dong, H.; Zhu, Y.; and Guibas, L. 2023. Sparsediff: Sparse-view feature distillation for one-shot dexterous manipulation. *arXiv preprint arXiv:2310.16838*.
- Wang, Y.; Xu, Z.; Huang, Y.; Wang, X.; Song, Z.; Gao, L.; Wang, C.; Tang, X.; Zhao, Y.; Cohan, A.; et al. 2025. DyFlow: Dynamic Workflow Framework for Agentic Reasoning. *arXiv preprint arXiv:2509.26062*.
- Xie, Y.; Zhou, C.; Gao, L.; Wu, J.; Li, X.; Zhou, H.-Y.; Liu, S.; Xing, L.; Zou, J.; Xie, C.; and Zhou, Y. 2025. MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine. *arXiv:2408.02900*.
- Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Liu, Z.; Ding, S.; Wu, S.; Ma, Y.; Duan, H.; Zhang, W.; et al. 2025. A Simple Yet Effective Multi-Modal Reward Model. *arXiv preprint arXiv:2501.12368*.
- Zhang, W.; Wang, M.; Liu, G.; Huixin, X.; Jiang, Y.; Shen, Y.; Hou, G.; Zheng, Z.; Zhang, H.; Li, X.; et al. 2025. Embodied-Reasoner: Synergizing Visual Search, Reasoning, and Action for Embodied Interactive Tasks. *arXiv preprint arXiv:2503.21696*.
- Zhao, B.; Wang, Z.; Fang, J.; Gao, C.; Man, F.; Cui, J.; Wang, X.; Chen, X.; Li, Y.; and Zhu, W. 2025. Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning in Foundation Models via Reinforcement Learning. *arXiv preprint arXiv:2504.12680*.
- Zhou, J.; He, X.; Sun, L.; Xu, J.; Chen, X.; Chu, Y.; Zhou, L.; Liao, X.; Zhang, B.; Afvari, S.; et al. 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications*, 15(1): 5649.