# VLASER: VISION-LANGUAGE-ACTION MODEL WITH SYNERGISTIC EMBODIED REASONING

**Ganlin Yang**[1,2*], **Tianyi Zhang**[4,2*], **Haoran Hao**[5,2*], **Weiyun Wang**[6,2], **Yibin Liu**[9,3], **Dehui Wang**[3]
**Guanzhou Chen**[3,2], **Zijian Cai**[10,3], **Junting Chen**[8,2], **Weijie Su**[2], **Wengang Zhou**[1], **Yu Qiao**[2]
**Jifeng Dai**[7,2], **Jiangmiao Pang**[2], **Gen Luo**[2], **Wenhai Wang**[2], **Yao Mu**[3,2†], **Zhi Hou**[2†]

[1]University of Science and Technology of China [2]Shanghai AI Laboratory
[3]Shanghai Jiao Tong University [4]Zhejiang University [5]Nanjing University [6]Fudan University
[7]Tsinghua University [8]NUS [9]Northeastern University [10]Shenzhen University

Project Page: Vlaser

## ABSTRACT

While significant research has focused on developing embodied reasoning capabilities using Vision-Language Models (VLMs) or integrating advanced VLMs into Vision-Language-Action (VLA) models for end-to-end robot control, few studies directly address the critical gap between upstream VLM-based reasoning and downstream VLA policy learning. In this work, we take an initial step toward bridging embodied reasoning with VLA policy learning by introducing **Vlaser** – a **V**ision-**L**anguage-**A**ction Model with **s**ynergistic **e**mbodied **r**easoning capability, which is a foundational vision-language model designed to integrate high-level reasoning with low-level control for embodied agents. Built upon the high-quality Vlaser-6M dataset, Vlaser achieves state-of-the-art performance across a range of embodied reasoning benchmarks—including spatial reasoning, embodied grounding, embodied QA, and task planning. Furthermore, we systematically examine how different VLM initializations affect supervised VLA fine-tuning, offering novel insights into mitigating the domain shift between internet-scale pre-training data and embodied-specific policy learning data. Based on these insights, our approach achieves state-of-the-art results on the WidowX benchmark and competitive performance on the Google Robot benchmark. The code, model and data are available at `https://github.com/OpenGVLab/Vlaser/`.

## 1 INTRODUCTION

Embodied artificial intelligence (AI) (Chrisley, 2003) aims to endow agents with the ability to perceive, understand, and act in the physical world. Achieving such intelligence requires not only accurate perception and language understanding but also embodied reasoning and effective control, which together define the paradigm of vision-language-action (VLA) models. Developing foundation models that possess strong reasoning and control capabilities is therefore an important advancement toward general-purpose embodied AI.

In this context, vision-language models (VLMs) (OpenAI, 2023; Liu et al., 2023; Chen et al., 2024; Bai et al., 2025; Team et al., 2023) emerge as natural candidates to enhance embodied agents in perception generalization and reasoning ability. Following this paradigm, extensive embodied vision-language models (Azzolini et al., 2025; Team et al., 2025c) emerge from enhancing the key ability for an embodied agent in grounding, planning, and spatial reasoning. Meanwhile, a significant body of work extends vision-language models (VLMs) into vision-language-action models (VLAs) (Kim et al., 2024; Intelligence et al., 2025; Driess et al., 2025) for robot control. While there are some approaches (Intelligence et al., 2025; Driess et al., 2025) that demonstrate the effectiveness of co-training with web data for the generalization in robot manipulation, it remains poorly understood which multi-modal data streams/abilities are most critical for improving downstream VLA models. In this paper, we aim to construct Vlaser, an embodied vision-language model that possesses strong

*Equal contribution. †Corresponding authors.

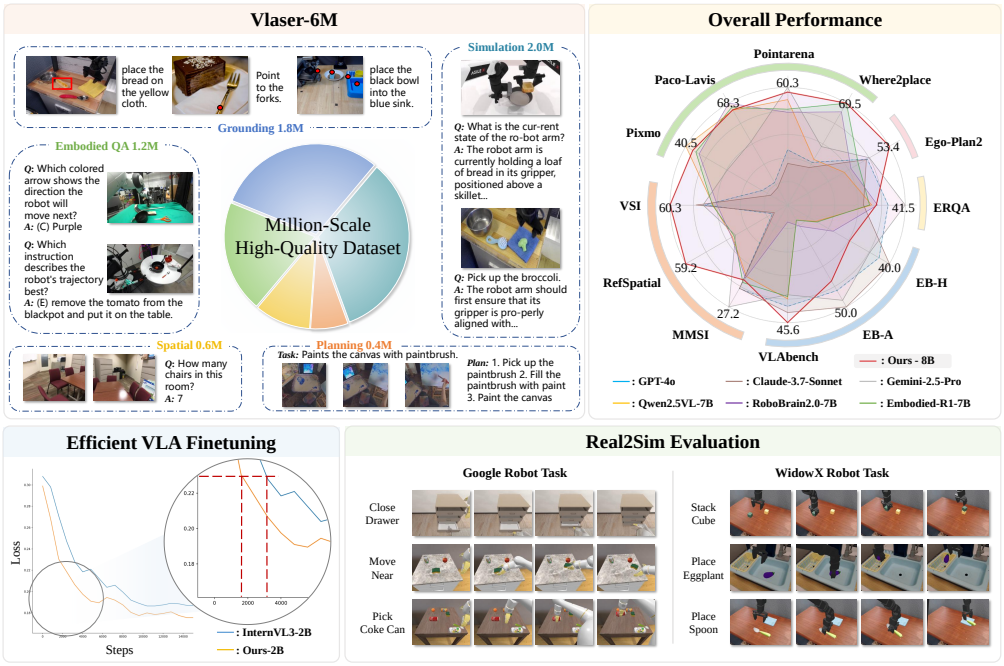



Figure 1: **Overall framework, capabilities, and evaluation of Vlaser. Top-left:** Composition of the Vlaser-6M dataset, featuring multi-task embodied data—including QA, grounding, spatial reasoning, and planning—along with in-domain simulation-sourced pairs. **Top-right:** A LiDAR visualization illustrating the state-of-the-art embodied reasoning capability of the Vlaser VLM. **Bottom-left:** The pre-trained Vlaser VLM significantly accelerates convergence in downstream Vision-Language Action model (VLA) policy learning on WidowX platform (Walke et al., 2023a). **Bottom-right:** Successful closed-loop operation of an agent powered by Vlaser within the SimplerEnv benchmark (Li et al., 2024c).

embodied reasoning capabilities, and subsequently answer this question based on the corresponding vision-language-action models.

Despite advancements in vision-language models (Chen et al., 2024; Bai et al., 2025), the capabilities of operating as an embodied agent remain severely constrained. In particular, navigation and traditional manipulation approaches rely heavily on planning-based control (Huang et al., 2022; Gasparetto et al., 2015; Zhang et al., 2018), which requires a strong foundational ability in grounding and planning. Planning and Grounding are cornerstones of the agents embodied in the physical world. Meanwhile, spatial understanding increasingly attracts the interest of the community in addressing the spatial perception ability of VLM. To this end, we firstly aim to introduce an embodied vision-language model specifically enhanced for the aboved embodied reasoning capabilities. Specifically, we construct the Vlaser data engine, which enables the systematic construction of the Vlaser-6M dataset by curating, reorganizing, and annotating public datasets from the Internet. As illustrated in Figure 1, the resulting dataset spans a wide spectrum of embodied reasoning tasks–including general embodied QA, visual grounding, spatial intelligence, task planning, and in-domain simulation data. Leveraging this comprehensive data foundation, Vlaser achieves state-of-the-art performance across a variety of embodied reasoning benchmarks, demonstrating strong generalization in both open-loop inference and closed-loop control settings.

Existing Vision-Language-Action (VLA) models (Black et al., 2024; Cheng et al., 2024; Kim et al., 2024; Intelligence et al., 2025) typically fine-tune pre-trained Vision-Language Models (VLMs) for robot control. However, the selection of an optimal VLM backbone – one that accelerates convergence and improves success rates when used as initialization for end-to-end VLA policy learning, remains an under-explored research problem. To address this gap, we systematically investigate the VLM-to-VLA adaptation paradigm using our enhanced embodied vision-language model and associated data engine. Our experiments reveal an important insight: although out-of-domain embodied reasoning

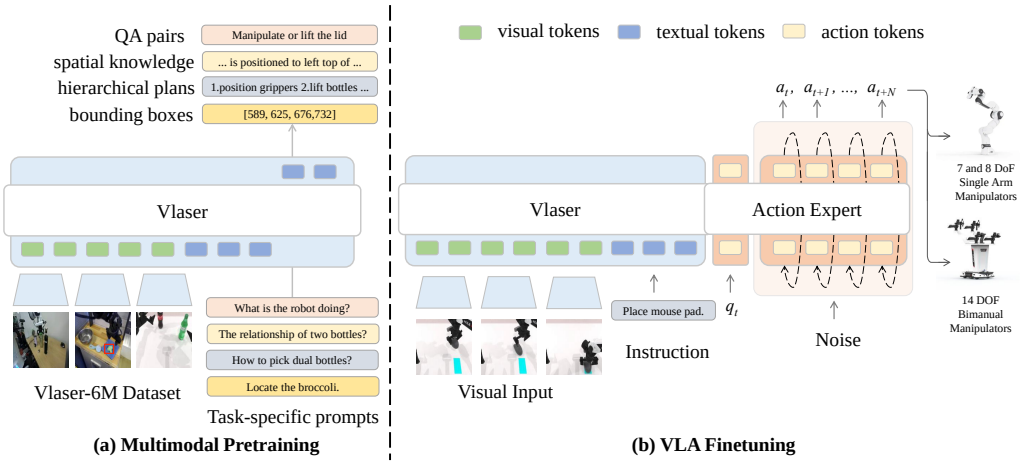



Figure 2: **An illustration of Vlaser architecture.** Vlaser includes two components and corresponding training phases: 1) the Multimodal Pretraining is for embodied reasoning enhancement based on the corresponding data engine; 2) VLA training is performed on the action expert module, which handles low-level control based on flow matching action generation.

data significantly improve upstream reasoning capabilities as measured by standard benchmarks, these gains may not translate directly or prominently to downstream VLA performance. In contrast, in-domain data – annotated directly on robot interaction datasets such as Open X-Embodiment (O'Neill et al., 2024) proves substantially more effective in accelerating convergence and increasing task success rates during VLA fine-tuning. We believe this observation provides significant insights for future embodied vision-language model construction: It is urgent to shrink the domain gap between current embodied perception and reasoning benchmarks to the real-world embodied tasks, and thus facilitate the closed-loop evaluation for the corresponding robot embodiment.

In summary, the principal contributions of Vlaser are as follows.

**An open-source embodied vision-language model and dataset.** We introduce Vlaser, an adaptable vision-language model that enhances InternVL with embodied reasoning capabilities and end-to-end robot control. The full model weights, modular data generation pipeline, training and evaluation code, and the accompanying Vlaser-6M dataset will be made publicly available to support reproducibility and future research.

**Systematic analysis of data effectiveness for VLA transfer.** We conduct a thorough investigation into which types of vision-language pretraining data contribute most effectively to downstream Vision-Language-Action (VLA) policy learning. Our findings offer practical insights for constructing task-aware data streams that bridge the gap between Internet-scale pretraining and embodied-specific fine-tuning.

**State-of-the-art performance across embodied benchmarks.** Among models of comparable scale, Vlaser achieves top-tier results on a comprehensive set of embodied reasoning benchmarks—spanning visual grounding, task planning, spatial reasoning, and simulation-based robot evaluation, demonstrating its strong generalization and applicability to both open-loop inference and closed-loop control scenarios.

## 2 METHOD

Vlaser aims to integrate embodied reasoning with end-to-end robot control for embodied agents, and identify the most crucial VLM data stream for VLA models. We first present the Vlaser structures in Section 2.1. Then, we illustrate the data engine in Section 2.2. Section 2.3 discusses the training recipe that includes embodied reasoning pretraining and vision-language-action finetuning.

### 2.1 MODEL STRUCTURE

The structure of Vlaser consists of two major components: the typical vision-language backbone (Chen et al., 2024; Liu et al., 2023) and the action expert for low-level control, as shown in Figure 2. We illustrate the two components respectively in this section.

**VLM Backbone** Vision-language models (VLMs) are key candidates for embodied agents, providing both perception and reasoning abilities. Vlaser , built on InternVL3 (Zhu et al., 2025), integrates embodied reasoning with robot control for embodied agents. While InternVL3 excels in multimodal and linguistic tasks across various model sizes, Vlaser focuses on two sizes—2B and 8B—optimized for the computational constraints of robots. These models utilize InternViT (Chen et al., 2024) as the vision encoder, paired with Qwen2.5-1.5B and Qwen2.5-7B LLMs (Qwen et al., 2025). Unlike typical multimodal MLLMs, Vlaser emphasizes embodied common-sense reasoning and end-to-end robot control capabilities.

**Action Expert** There are a large number of MLLMs (Team et al., 2025a; NVIDIA et al., 2025a) that enhance the ability of embodied common-sense reasoning for agents, while a few approaches equip the embodied MLLMs with end-to-end robot control. Vlaser extends the MLLMs with a low-level robot control and verifies the capability of different data streams in downstream VLA finetuning. Following (Intelligence et al., 2025), we design an action expert based on the open-source vision-language model (Chen et al., 2024; Zhu et al., 2025). Meanwhile, we utilize the flow matching (Lipman et al., 2023a) for action prediction based on the llava-like vision-language structure, while sharing the self-attention among the language model and action expert module. Specifically, we encode the robot state as a state token and noised actions as action tokens, and input them into the action expert. Meanwhile, we utilize non-causal attention for the VLA stream. During inference, we denoise the actions based on the image observation, language instruction, as well as the current robot state.

### 2.2 VLASER DATA ENGINE

This section outlines the composition of the Vlaser-6M data engine, a cornerstone for the model's embodied reasoning capabilities. Here we present the overall data scale and sources for each reasoning modality, while more details about the construction methodologies are provided in Appendix A.2.

**Embodied Grounding Data** The Vlaser dataset incorporates two distinct 2D grounding formats—bounding boxes and center points—both normalized to the range [0, 1000] to ensure consistent and resolution-invariant grounding predictions across diverse image resolutions. Specifically, we collect 1.5 million high-quality question-answer pairs that support multiple grounding tasks: predicting bounding boxes from open-vocabulary descriptions, localizing object center points based on textual descriptions, and identifying objects from given spatial coordinates. The data is sourced from several open embodied grounding datasets, including RoboPoint (Yuan et al., 2024), ShareRobot (Ji et al., 2025), Pixmo-Points (Deitke et al., 2025), Paco-LaVIS (Ramanathan et al., 2023), and RefSpatial (Zhou et al., 2025a). To further enhance generalization capabilities for open-world and open-vocabulary scenarios, we also generate an additional 300k point and bounding box annotations derived from segmentation masks in the SA-1B dataset (Kirillov et al., 2023). This combination of curated human annotations and synthetically enriched data aims to bolster both the diversity and scalability of visual grounding under real-world embodied settings.

**General and Spatial Reasoning Data** The Vlaser dataset integrates 1.2 million question-answer pairs dedicated to general Robotic Visual Question Answering (RoboVQA) tasks, along with an additional 500k data items specifically designed to enhance spatial intelligence. This comprehensive data composition substantially strengthens the model's capabilities in general state perception and 3D spatial reasoning. For the RoboVQA component, data is aggregated from multiple established sources, including RoboVQA (Sermanet et al., 2024), Robo2VLM (Chen et al., 2025b), RoboPoint (Yuan et al., 2024), RefSpatial (Zhou et al., 2025a), OWMM-Agent (Chen et al., 2025a), among others. To support spatial understanding and reasoning, we incorporate open-source datasets such as SPAR (Zhang et al., 2025), SpaceR-151k (Ouyang et al., 2025), and VILASR (Wu et al., 2025). Furthermore, we augment these with 100k manually annotated spatial understanding samples generated from publicly available 3D scene datasets—including ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al.,

2023), CA-1M (Lazarow et al., 2025), and ARKitScenes (Baruch et al., 2021). The integration of these diverse and high-quality data sources effectively enhances the model's spatial awareness and supports more robust performance in complex embodied reasoning tasks.

**Planning Data** To tackle complex tasks, it is essential to decompose them into manageable sub-tasks and solve them step by step. This capability is commonly referred to as planning. Effective planning allows robots to combine basic skills and generalize to new scenarios. We collected 400k training data to strengthen the model's planning ability, encompassing both language-based planning data and multimodal tasks. These include Alpaca-15k-Instruction (Wu et al., 2023) and MuEP (Li et al., 2024a). To further enhance environmental understanding and reasoning for complex decision-making, we incorporated training data with detailed reasoning processes from WAP (Shi et al., 2025). To improve the model's ability to comprehend complex instructions and execute tasks, we followed the annotations of LLaRP (Szot et al., 2024) to initialize planning tasks in Habitat (Szot et al., 2021) and generate planning trajectories to accomplish these tasks. In addition, we integrated egocentric video datasets such as EgoPlan-IT (Chen et al., 2023) and EgoCOT (Mu et al., 2023), which closely align with the observational perspective of embodied agents and provide valuable planning examples.

**In-Domain Data for downstream VLAs** To facilitate the end-to-end policy learning for Vision-Language Action Models (VLAs), we further generate 2 million in-domain multimodal question-answer pairs tailored for VLM pretraining. These data are specifically designed to align with the embodied reasoning context and enhance the model's ability to perceive, reason, and plan in interactive environments. The in-domain data is sourced from simulation platforms SimplerEnv (Li et al., 2024d) and RoboTwin (Chen et al., 2025c). Within SimplerEnv, data is generated for two distinct robotic embodiments: the Google Robot (Brohan et al., 2023b;a; O'Neill et al., 2024) and the WidowX Robot (Walke et al., 2023a), and within RoboTwin, data is generated from dual-arm Aloha-AgileX Robot. The question-answer pairs encompass the specialized categories including embodied grounding, spatial intelligence, planning and general VQA for robot states as described above. The detailed methodology for constructing and filtering each of the in-domain data in simulation is described in Appendix A.2.

### 2.3 Training Recipe

Vlaser adopts a two-stage training recipe, designed to optimize both embodied reasoning and robot control. It includes a VLM pretraining followed by a VLA finetuning. In this section, we elaborate on the training recipe among all phrases.

**Vision-Language Pretraining** Vlaser is developed by supervised fine-tuning (SFT) InternVL3 (Zhu et al., 2025) on embodied-related datasets, including those focused on grounding, planning, and spatial intelligence. In the first training phase, we fine-tune InternVL3 using auto-regressive language modeling loss. In particular, given the input images $x \in \mathbb{R}^{t\times h\times w\times 3}$ and textual prompt $y \in \mathbb{R}^{l}$, the language modeling loss $\mathcal{L}_{lm}$ can be defined by

$$\mathcal{L}_{lm} = -\log p(t_N|\mathcal{F}_{\text{v}}(x;\theta_v), \mathcal{F}_{\text{t}}(y), t_{0:N-1}; \Theta), \tag{1}$$

where $p \in \mathbb{R}^m$ is the next-token probability and $m$ denotes the vocabulary size. Here, $\mathcal{F}_{\text{v}}(\cdot)$ denotes the ViT and the MLP, and $\theta_v$ is their parameters. $\mathcal{F}_{\text{t}}(\cdot)$ is the textual tokenizer. $\Theta$ are the parameters of the LLM. $t_i$ denotes the *i-th* predicted word.

**Vision-Language-Action Finetuning** For robot policy learning, we optimize the model using an additional incorporated action expert module trained on robot-specific datasets. The action expert is analogous to a mixture of experts(MoE) (Shazeer et al., 2017b; Du et al., 2022; Zhou et al., 2024) architecture with two mixture elements, while the original part of parameter is used for image and text inputs, and the additionally separate set of weights for the robotics-specific (action and state) tokens inputs and outputs are referred as the *action expert*. Vlaser integrates a flow-matching-based action expert to predict a sequence of future actions from a single-frame observation. Specifically, we denote the action chunk $\mathbf{A}_t = [\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+H-1}]$, where $\mathbf{a}_t$ represents the action in the current timestep $t$ and $H$ represents the action horizon. $\mathbf{o}_t = [\mathbf{I}_t^1, ..., \mathbf{I}_t^n, l_t, \mathbf{q}_t]$ indicates the observations (image $\mathbf{I}_t^i$ with $n$ views, language $l_t$ and robot state $\mathbf{q}_t$) at action timestep $t$. $\mathbf{I}_t^i$, $l_t$ and $\mathbf{q}_t$ are encoded via corresponding encoders and then projected via a linear projection layer into the same embedding space. $\theta$ represents the action expert network and $\tau \in [0, 1]$ represents the flow matching timesteps.

Table 1: **Comparison with existing close-sourced, open-sourced and embodied-related VLMs on 12 general embodied reasoning benchmarks, spanning from embodied QA, planning, embodied grounding to spatial intelligence and close-loop simulation evaluation.** Avg denotes the normalized average performance of all the benchmarks. The best, second best and third best score among all the baselines are colored in red, orange and yellow.

| Model | QA | Planning | Embodied Grounding | | | | Spatial Intelligence | | | Simulation | | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ERQA** | **Ego-Plan2** | **Where2place** | **Pointarena** | **Paco-Lavis** | **Pixmo-Points** | **VSIBench** | **RefSpatial** | **MMSIBench** | **VLABench** | **EB-ALFRED** | **EB-Habitat** | |
| ▼ *Closed-source MLLMs:* | | | | | | | | | | | | | |
| GPT-4o-20241120 | 47.0 | 41.8 | 29.1 | 29.5 | 16.2 | 10.8 | 42.5 | 8.8 | 30.3 | 39.3 | 56.3 | 59.0 | 34.2 |
| Claude-3.7-Sonnet | 35.5 | 41.3 | 25.6 | 22.2 | 12.4 | 7.2 | 47.0 | 7.7 | 30.2 | 41.7 | 67.0 | 65.7 | 33.6 |
| Gemini-2.5-Pro | 55.0 | 42.9 | 39.9 | 62.8 | 45.5 | 25.8 | 43.4 | 30.3 | 36.9 | 34.8 | 62.7 | 53.0 | 44.4 |
| ▼ *Small Size MLLMs:* | | | | | | | | | | | | | |
| ChatVLA-2B | 34.3 | 25.3 | 3.7 | 10.1 | 10.2 | 2.1 | 2.4 | 0.9 | 20.1 | 0.0 | 0.0 | 0.0 | 9.1 |
| InternVL3-2B | 31.5 | 30.9 | 5.2 | 7.1 | 15.4 | 1.4 | 31.5 | 1.8 | 25.3 | 19.4 | 1.3 | 12.0 | 15.2 |
| Qwen2.5VL-3B | 35.3 | 30.3 | 31.0 | 41.7 | 67.4 | 36.6 | 27.9 | 24.9 | 26.5 | 31.3 | 6.7 | 19.7 | 31.6 |
| Embodied-R1-3B | 36.0 | 36.0 | 35.1 | 45.3 | 68.3 | 36.6 | 28.0 | 28.5 | 26.0 | 24.6 | 7.0 | 19.3 | 32.5 |
| RoboBrain2.0-3B | 37.3 | 41.8 | 64.2 | 46.0 | 67.6 | 36.9 | 28.8 | 46.5 | 26.8 | 18.1 | 0.0 | 10.0 | 35.3 |
| Vlaser-2B | 35.8 | 38.3 | 74.0 | 57.8 | 72.5 | 44.6 | 57.5 | 43.0 | 23.6 | 23.1 | 42.3 | 30.7 | 45.3 |
| ▼ *Medium Size MLLMs:* | | | | | | | | | | | | | |
| Magma-8B | 29.3 | 27.9 | 10.9 | 29.6 | 15.3 | 10.1 | 12.7 | 4.5 | 26.2 | 8.5 | 0.0 | 0.0 | 14.6 |
| Cosmos-Reason1-7B | 39.3 | 26.9 | 11.4 | 40.8 | 61.8 | 23.6 | 33.9 | 5.4 | 26.4 | 35.5 | 4.0 | 5.3 | 26.2 |
| VeBrain-7B | 38.3 | 27.3 | 33.1 | 38.9 | 55.1 | 20.1 | 39.9 | 20.6 | 28.3 | 25.9 | 5.7 | 12.3 | 28.8 |
| InternVL3-8B | 35.3 | 40.0 | 10.0 | 14.2 | 21.1 | 5.7 | 42.1 | 5.6 | 25.7 | 24.7 | 19.0 | 23.7 | 22.3 |
| Qwen2.5VL-7B | 39.3 | 29.7 | 31.1 | 56.3 | 68.0 | 43.5 | 38.2 | 32.1 | 25.9 | 36.4 | 10.0 | 18.3 | 35.7 |
| Embodied-R1-7B | 38.3 | 37.1 | 69.5 | 51.2 | 69.9 | 39.2 | 38.6 | 31.1 | 28.1 | 35.5 | 10.0 | 19.0 | 38.9 |
| RoboBrain2.0-7B | 42.0 | 33.2 | 63.6 | 49.5 | 73.1 | 37.8 | 36.1 | 32.5 | 26.5 | 6.6 | 14.0 | 29.3 | 37.0 |
| Vlaser-8B | 41.0 | 53.4 | 69.5 | 60.3 | 68.3 | 40.5 | 60.3 | 59.2 | 27.2 | 45.6 | 50.0 | 40.0 | 51.3 |

For the action chunk, $\mathbf{A}_t^\tau = \tau \mathbf{A}_t + (1-\tau)\epsilon$ is the corresponding noisy action chunk, and we train the network to output $\mathbf{v}_\theta(\mathbf{A}_t^\tau, \mathbf{o}_t)$ to match the denoising vector field $\mathbf{u}(\mathbf{A}_t^\tau | \mathbf{A}_t) = \epsilon - \mathbf{A}_t$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Therefore, the VLA optimization loss is as follows,

$$\mathcal{L}_{vla} = \mathbb{E}_{p(\mathbf{A}_t|\mathbf{o}_t)} \left\| \mathbf{v}_\theta(A_t^\tau, \mathbf{o}_t) - \mathbf{u}(\mathbf{A}_t^\tau | \mathbf{A}_t) \right\|^2 \tag{2}$$

Formally, following prior flow-matching based VLA works (Black et al., 2024; Zren, 2025), We sample the action chunks from the robot episodes and flow-matching timesteps to optimize the network. At inference, we generate actions by integrating the learned vector field from $\tau = 0$ to $\tau = 1$, starting with random noise $\mathbf{A}_t^0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, as follows,

$$\mathbf{A}_t^{\tau+\delta} = \mathbf{A}_t^\tau + \delta \mathbf{v}_\theta(\mathbf{A}_t^\tau, \mathbf{o}_t) \tag{3}$$

where $\delta$ is the integration step size. In our experiments, we set $H$ as 4, and $\delta$ as $0.1(\delta^{-1} = 10$ integration steps) at inference time for the improvement of inference efficiency. We aim to identify the most effective VLMs for downstream VLA fine-tuning and bridge the gap between foundational VLMs and their performance in downstream VLA tasks, thus shedding light on the future construction of embodied VLMs. Currently, the SimplerEnv benchmark, including Bridge (Walke et al., 2023b) and Google Robot (Jang et al., 2022; Brohan et al., 2023b) datasets, provides numerous training episodes (Over 5M images) and corresponding Real-to-Sim benchmarks. We thus majorly analyze the most effective data stream for VLA finetuning based on SimplerEnv.

## 3 Experiments

### 3.1 Performance on Embodied Reasoning Capability

**Evaluation Datasets** We conduct a comprehensive evaluation of embodied reasoning capabilities across a total of 12 benchmarks, covering a wide spectrum of tasks including embodied question answering, task planning, embodied grounding, spatial intelligence, and closed-loop simulation evaluation. The evaluated benchmarks consist of: ERQA (Team et al., 2025b), Ego-Plan2 (Qiu et al., 2024), Where2place (Yuan et al., 2024), Pointarena (Cheng et al., 2025), Paco-Lavis (Ramanathan et al., 2023), Pixmo-Points (Deitke et al., 2025), VSI-Bench (Yang et al., 2025b), RefSpatial-Bench (Zhou et al., 2025a), MMSI-Bench (Yang et al., 2025d), VLABench (Zhang et al., 2024), and EmbodiedBench (Yang et al., 2025c). For EmbodiedBench, we further assess performance in two simulation environments ALFRED (Shridhar et al., 2020) and Habitat (Szot et al., 2021).

Table 2: **SimplerEnv Evaluation on WidowX Robot Tasks.** Avg indicates the average success rate among the four tasks. Model sizes are indicated within parentheses. The result of RT-1-X (Brohan et al., 2023b), Octo-Base (Team et al., 2024), OpenVLA (Kim et al., 2024), RoboVLM (Liu et al., 2025) and SpatialVLA (Qu et al., 2025b) are from (Qu et al., 2025b) while the results of $\pi_0$ (Black et al., 2024) is from (Zren, 2025).

| Model | Carrot on plate | Put eggplant in basket | Spoon on towel | Stack Cube | Avg |
|---|---|---|---|---|---|
| RT-1-X (35M) (Brohan et al., 2023b) | 4.2% | 0% | 0% | 0% | 1.1% |
| Octo-Base (93M) (Team et al., 2024) | 8.3% | 43.1% | 12.5% | 31.9% | 16.0% |
| OpenVLA (7B) (OpenAI, 2023) | 0% | 4.1% | 0% | 0% | 1.0% |
| RoboVLM (2B) (Liu et al., 2025) | 25.0% | 58.3% | 29.2% | 12.5% | 31.3% |
| SpatialVLA (4B) (Qu et al., 2025b) | 25.0% | 100.0% | 16.7% | 62.5% | 42.7% |
| $\pi_0$ (3B) (Black et al., 2024) | 55.8% | 79.2% | 63.3% | 21.3% | 54.9% |
| InternVL3-2B | 42.9% | 57.1% | 55.8% | 11.3% | 41.8% |
| Vlaser-OOD (2B) | 60.8% | 35.4% | 56.7% | 20.0% | 43.2% |
| Vlaser-QA (2B) | 55.8% | 83.3% | 77.9% | 33.3% | 62.6% |
| Vlaser-Spatial (2B) | 48.3% | 81.7% | 76.7% | 36.7% | 60.8% |
| Vlaser-Grounding (2B) | 47.5% | 80.8% | 80.0% | 39.6% | 62.0% |
| Vlaser-All (2B) | 52.5% | 87.9% | 76.6% | 43.3% | 65.1% |

**Baselines** Since our method, Vlaser is trained at two model scales – 2B and 8B parameters, we categorize the compared baseline methods into three groups for a systematic evaluation: 1) **State-of-the-art closed-source models**, including GPT-4o (OpenAI, 2025), Claude-3.7-Sonnet (Anthropic, 2025), and Gemini-2.5-Pro (Comanici et al., 2025); 2) **Small-scale MLLMs (2B – 3B parameters)**, comprising ChatVLA-2B (Zhou et al., 2025b), InternVL3-2B (Zhu et al., 2025), Qwen2.5-VL-3B (Bai et al., 2025), Embodied-R1-3B (Yuan et al., 2025), and RoboBrain2.0-3B (Team et al., 2025a); 3) **Medium-scale MLLMs (7B – 8B parameters)**, including Magma-8B (Yang et al., 2025a), Cosmos-Reason1-7B (NVIDIA et al., 2025a), VeBrain-7B (Luo et al., 2025), InternVL3-8B (Zhu et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), Embodied-R1-7B (Yuan et al., 2025), and RoboBrain2.0-7B (Team et al., 2025a).

The overall experimental results are presented in Table 1. As shown in Table 1, compared to the base models InternVL3-2B and InternVL3-8B used as initialization for our supervised finetuning, our Vlaser yields substantial improvements across all embodied reasoning capabilities, with particularly notable gains in embodied grounding and simulation-based evaluation. For example, the average score increases from 15.2 to 45.3 for the 2B model, and from 22.3 to 51.3 for the 8B model. *These significant performance gains underscore the high quality and effectiveness of the Vlaser-6M dataset in enhancing embodied reasoning abilities.* An interesting observation emerges that when finetuning on the same Vlaser-6M dataset, a smaller sized Vlaser-2B outperforms Vlaser-8B on simple point grounding tasks that require direct, short answers. Conversely, Vlaser-8B demonstrates superior performance on more complex tasks such as multi-step planning and closed-loop simulation evaluation, which often benefit from chain-of-thought (CoT) reasoning. This scaling behavior indicates the importance of appropriate model size selection based on target application requirements.

When compared against current state-of-the-art embodied-specific VLMs, including RoboBrain2.0 (Team et al., 2025a) and Embodied-R1 (Yuan et al., 2025), our method, Vlaser still achieves superior performance on the majority of benchmarks while remaining highly competitive on the remainder, **ultimately attaining the highest overall score** (by +10% margin overall). These results indicate that Vlaser delivers a well-balanced and robust capability set, performing strongly across multiple dimensions of embodied intelligence – from embodied question answering and state estimation to future action planning, visual grounding, spatial reasoning, and closed-loop simulation. Such comprehensive competence highlights its suitability as a versatile backbone for embodied AI brains. In the following section, we further examine how these enhanced reasoning capabilities, embedded within VLMs, translate into improved performance when fine-tuned for downstream Vision-Language Action models (VLAs) in simulation manipulation scenarios.

### 3.2 Performance on downstream Close-Loop Robot Tasks

**Finetuning Datasets** We firstly conduct extensive experiments on SimplerENV (Li et al., 2024d) to evaluate the performance of Vlaser and Vlaser data engine on closed-loop robotic manipulation

Table 3: **Comparison with existing methods in SimplerEnv on Google Robot tasks.** Avg indicates the average success rate among the three tasks. Model sizes are indicated within parentheses. The results of TraceVLA (Zheng et al., 2024), RT-1-X (Brohan et al., 2023b), Octo-Base (Team et al., 2024), OpenVLA (Kim et al., 2024), RoboVLM (Liu et al., 2025), Emma-X (Sun et al., 2024), Magma (Yang et al., 2025a), GR00T N1.5(NVIDIA et al., 2025b) and $\pi_0$ (Black et al., 2024) are from (Lee et al., 2025).

| Model | Visual Matching | | | Avg | Variant Aggregation | | | Avg |
|---|---|---|---|---|---|---|---|---|
| | Pick Coke Can | Move Near | Drawer | | Pick Coke Can | Move Near | Drawer | |
| TraceVLA (7B) (Zheng et al., 2024) | 28.0% | 53.7% | 57.0% | 42.0% | 60.0% | 56.4% | 31.0% | 45.0% |
| RT-1-X (35M) (Brohan et al., 2023b) | 56.7% | 31.7% | 59.7% | 53.4% | 49.0% | 32.3% | 29.4% | 39.6% |
| Octo-Base (93M) (Team et al., 2024) | 17.0% | 4.2% | 22.7% | 16.8% | 0.6% | 3.1% | 1.1% | 1.1% |
| OpenVLA (7B) (Kim et al., 2024) | 16.3% | 46.2% | 35.6% | 27.7% | 54.5% | 47.7% | 17.7% | 39.8% |
| RoboVLM (2B) (Liu et al., 2025) | 77.3% | 61.7% | 43.5% | 63.4% | 75.6% | 60.0% | 10.6% | 51.3% |
| Emma-X (7B) (Sun et al., 2024) | 2.3% | 3.3% | 18.3% | 8.0% | 5.3% | 7.3% | 20.5% | 11.0% |
| Magma (8B) (Yang et al., 2025a) | 56.0% | 65.4% | 83.7% | 68.4% | 53.4% | 65.7% | 68.8% | 62.6% |
| GR00T N1.5 (2.1B) (NVIDIA et al., 2025b) | 69.3% | 68.7% | 35.8% | 52.4% | 46.7% | 62.9% | 17.5% | 43.7% |
| $\pi_0$ (3B) (Black et al., 2024) | 72.7% | 65.3% | 38.3% | 58.3% | 75.2% | 63.7% | 25.6% | 54.8% |
| InternVL3-2B | 94.3% | 78.8% | 19.0% | 64.0% | 80.4% | 72.7% | 11.1% | 54.7% |
| Vlaser-OOD(2B) | 85.0% | 76.3% | 44.9% | 68.7% | 74.4% | 69.2% | 10.3% | 51.3% |
| Vlaser-QA (2B) | 90.0% | 84.2% | 44.4% | 72.9% | 78.2% | 78.2% | 13.0% | 56.4% |
| Vlaser-Spatial (2B) | 83.0% | 77.9% | 56.0% | 72.3% | 77.7% | 73.2% | 13.2% | 54.7% |
| Vlaser-Grounding (2B) | 83.3% | 83.3% | 54.2% | 73.6% | 81.2% | 76.8% | 17.0% | 58.3% |
| Vlaser-All (2B) | 91.0% | 85.4% | 52.1% | 76.2% | 80.5% | 77.7% | 18.8% | 59.0% |

tasks. SimplerENV is an open-source suite of purpose-built simulated environments with nearly 150K episodes for evaluating real-world robot manipulation policies in a scalable, reproducible way. It targets the key real-to-sim gaps – control and vision so that simulated performance reliably tracks real-robot outcomes. Across Google Robot and WidowX/BridgeData V2 setups, SimplerEnv reports strong real-vs-sim correlations and faithfully reflects behavior under distribution shifts, enabling fast, comparable policy assessment without full digital twins. As a result, SimplerENV has been widely adopted for evaluating VLA models and has proven to reliably reflect the performance of the models on the real robot platform. To further demonstrate the generalizability of our method to other simulation platforms and embodiments, we also conduct experiments on RoboTwin (Mu et al., 2024; Chen et al., 2025c) platforms with Aloha-AgileX as bimanual embodiement. Robotwin is a scalable framework for bimanual manipulation, which integrates scalable training sets and pre-defined tasks as benchmarks for comprehensive robust bimanual manipulation.

Table 4: **Robotwin Evaluation on Aloha-AgileX Robot Tasks.** Avg indicates the average success rate among the 12 tasks. The results of RDT-1B (Liu et al., 2024) are from our self-implemented training for 30k steps, which aligns the training setting with Vlaser.

| Simulation Task | RDT-1B (Liu et al., 2024) | InternVL3-2B | Vlaser-OOD(2B) | Vlaser-QA (2B) | Vlaser-Spatial (2B) | Vlaser-Grounding (2B) | Vlaser-All (2B) |
|---|---|---|---|---|---|---|---|
| Beat block hammer | 28% | 12% | 20% | 18% | 20% | 32% | 40% |
| Click bell | 46% | 78% | 94% | 48% | 98% | 86% | 92% |
| Handover mic | 92% | 74% | 56% | 84% | 60% | 52% | 84% |
| Move can pot | 44% | 40% | 42% | 66% | 56% | 50% | 46% |
| Move pillbottles pad | 10% | 62% | 70% | 66% | 78% | 68% | 72% |
| Move playingcard away | 20% | 64% | 52% | 58% | 68% | 84% | 74% |
| Pick diverse bottles | 2% | 30% | 44% | 36% | 24% | 34% | 38% |
| Place burger fries | 42% | 36% | 46% | 88% | 82% | 46% | 42% |
| Place container plate | 82% | 72% | 70% | 84% | 78% | 82% | 84% |
| Place phone stand | 8% | 42% | 40% | 38% | 36% | 48% | 50% |
| Place mouse pad | 2% | 68% | 30% | 44% | 38% | 48% | 92% |
| Shake bottle | 66% | 92% | 90% | 98% | 96% | 98% | 96% |
| Avg. | 36.8% | 55.8% | 54.5% | 60.7% | 61.2% | 60.7% | 67.5% |

**Baselines** Alongside comparisons with other commonly used VLA models (Black et al., 2024; Kim et al., 2024; Liu et al., 2025), we conduct a clear self-comparable ablation study to evaluate the individual contributions of different in-domain data sources. Specifically, **InternVL3-2B** denotes the base InternVL model, while **Vlaser-OOD** refers to the Vlaser-2B model fine-tuned solely on Vlaser-6M out-of-domain (OOD) data specific for the embodied reasoning benchmarks in Sec. 3.1, without any in-domain data in Vlaser-6M dataset. Regarding the in-domain datas derived directly from the simulation platform, we categorize them into three types: embodied QA (including planning), embodied spatial intelligence, and embodied grounding. The corresponding fine-tuned Vlaser-2B models

Table 5: **Ablation Studies on WidowX Robot Tasks**

| Model | Predict Length | Execute Length | Sample Steps | Carrot on plate | Put eggplant in basket | Spoon on the towel | Stack cube | Avg |
|---|---|---|---|---|---|---|---|---|
| **InternVL-2B** | 4 | 4 | 10 | 42.9% | 57.1% | 55.8% | 11.3% | 41.8% |
| | 4 | 2 | 10 | 22.9% | 18.3% | 40.8% | 2.9% | 21.2% |
| | 2 | 2 | 10 | 34.6% | 22.9% | 54.2% | 2.9% | 28.7% |
| | 4 | 4 | 20 | 38.8% | 54.2% | 51.3% | 8.3% | 38.2% |
| **Vlaser-OOD(2B)** | 4 | 4 | 10 | 60.8% | 35.4% | 56.7% | 20.0% | 43.2% |
| | 4 | 2 | 10 | 50.0% | 21.7% | 30.0% | 12.1% | 28.5% |
| | 2 | 2 | 10 | 62.5% | 19.2% | 49.2% | 21.3% | 38.1% |
| | 4 | 4 | 20 | 57.5% | 29.2% | 54.6% | 17.1% | 39.6% |
| **Vlaser-QA(2B)** | 4 | 4 | 10 | 55.8% | 83.3% | 77.9% | 33.3% | 62.6% |
| | 4 | 2 | 10 | 44.2% | 64.2% | 59.6% | 36.3% | 51.1% |
| | 2 | 2 | 10 | 47.5% | 66.3% | 67.1% | 36.3% | 54.3% |
| | 4 | 4 | 20 | 56.3% | 85.0% | 76.7% | 35.0% | 63.3% |

are denoted as **Vlaser-QA**, **Vlaser-Spatial**, and **Vlaser-Grounding**, respectively. Additionally, we combine all the three in-domain data types to fine-tune a model referred to as **Vlaser-All**.

The full experimental results are presented in Table 2, Table 3 and Table 4. Notably, no clear improvement was observed when using Vlaser-OOD-2B as the initial backbone across all three benchmarks. The task success rates of **Vlaser-OOD** and the baseline **InternVL3-2B** remain close. Conversely, all models fine-tuned with in-domain data—**Vlaser-QA**, **Vlaser-Spatial**, **Vlaser-Grounding**, and **Vlaser-All**—exhibit significant performance gains, even though the model architecture and size remain unchanged. This observation illustrates the effectiveness of our Vlaser data engine, and meanwhile identifies that *there is no positive correlation between common embodied reasoning benchmarks and the performance of closed-loop control of the lower level for the specific embodiment of the robot*. We reckon it is the domain shift between the internet data and the corresponding robot embodiment (e.g., WidowX or Google Robot), and we find that the enhanced abilities in the same observation domain effectively facilitate the closed-loop success rate. Therefore, it is urgent to shrink the domain gap between the foundational models and real-world robot embodiment for closed-loop task completion.

Regarding the three types of in-domain data annotations, we experimentally find that incorporating any of them leads to significant performance gains over the baseline. We attribute these improvements primarily to the mitigation of visual observation domain shift. Furthermore, integrating all three types of in-domain data results in further performance enhancement. *This suggests that pretraining with diverse in-domain multimodal data, spanning general QA, grounding, and spatial intelligence, could best facilitates transfer learning for VLA policy learning and leads to improved task success rates.*

### 3.3 Ablation Studies

In this section, we adopt ablation studies regarding three key hyperparameters for VLA end-to-end training, *i.e.*,, the predicted action length $P$, the execute action length sizes $H$ as well as flow-matching sampling steps $\delta^{-1}$. By default we use $P$ as 4, with $H$ as 4, and $\delta^{-1}$ as 10. The results are shown in Table 5. It is clear to see that, no matter in which group of hyperparameter settings, the performance based on **Vlaser-OOD** shows slight improvement compared with the performance based on **InternVL3-2B**. Besides, there is a significant gains while using the model fine-tuned with in-domain data **Vlaser-QA**. This conclusion is as same as the results in 3.2, which demonstrates great robustness of our method.

## 4 Related Work

**Vision-Language Model for Embodied Reasoning** Enhancing the embodied reasoning capabilities of current state-of-the-art Vision-Language Models (VLMs) has emerged as a critical research direction. These capabilities encompass a range of competencies, including grounding (Yuan et al., 2024; Deitke et al., 2025; Cheng et al., 2025) which identifies affordances that enable embodied agents to perform manipulations, spatial intelligence (Yang et al., 2025b;d), such as object counting and spatial relationship understanding, as well as task planning (Chen et al., 2023; Qiu et al., 2024), which involves assessing the current state and determining subsequent actions to be executed. Gemini Robotics-ER (Team et al., 2025b) integrates embodied reasoning into its core visual-language model (VLM), in parallel, a number of data-driven methodologies have emerged to support such reasoning

capabilities. For instance, Cosmos-Reason1 (NVIDIA et al., 2025a), VeBrain (Luo et al., 2025), MolmoAct (Lee et al., 2025), and EmbodiedOneVision (Qu et al., 2025a) each contribute curated datasets specifically designed for embodied reasoning tasks, emphasizing aspects such as multi-modal instruction following and action-aware visual-language alignment. Furthermore, several frameworks – including the RoboBrain series (Ji et al., 2025; Team et al., 2025a), Embodied R1 (Yuan et al., 2025), and Robix (Fang et al., 2025) incorporate Reinforcement Fine-Tuning (RFT) and synthesize spatio-temporal reasoning datasets enriched with structured thought traces. These approaches aim to enhance models' capacity for causal reasoning and long-horizon task decomposition. Distinguished from these prior efforts, our work not only achieves competitive, and in some cases superior performance on established embodied reasoning benchmarks, but also provides an in-depth analysis of the synergistic relationship between pre-trained VLMs and downstream Vision-Language Action Models (VLAs), offering insights that bridge model capabilities and real-world deployment.

**Vision-Language-Action models.** Developing a generalist model remains a central challenge in robotics. Inspired by the strong generalization abilities of vision-language models (VLMs) (OpenAI, 2023; Chen et al., 2024; Bai et al., 2025; Team et al., 2023) trained on large-scale internet data, researchers have proposed vision-language-action (VLA) models, which have demonstrated promising performance (Brohan et al., 2023a; Kim et al., 2024; Qu et al., 2025b; Hou et al., 2025). Compared to traditional robot policies, VLA models are pretrained on large-scale robotics datasets and exhibit improved generalization across object categories and visual observations. Building on recent progress, researchers have incorporated techniques such as diffusion (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023), flow matching (Lipman et al., 2023b), and mixture-of-experts (MoE) (Shazeer et al., 2017a) into VLA models, and have adopted larger, more capable VLMs as their backbones. These advances have enabled VLA models to tackle a wider range of complex real-world manipulation tasks. Efforts have been made to enhance specific reasoning abilities in embodied scenarios (Team et al., 2025b; Ji et al., 2025; NVIDIA et al., 2025a). In parallel, several studies (Intelligence et al., 2025; Driess et al., 2025; Zhou et al., 2025b) explore unified training frameworks for VLMs and VLAs to leverage the reasoning capacity of VLMs. However, the relationship between high-level multimodal reasoning and low-level control performance remains largely unexplored. It is still unclear which specific multimodal abilities such as spatial understanding, grounding, or planning and which types of training data most effectively enhance the control capabilities of a VLA model. In this work, we take an initial step toward analyzing this relationship by a systematic evaluation, and also propose the latest foundational model with strong embodied multimodal understanding and action prediction.

## 5 Limitations

One limitation of this work is the absence of real-robot experiments. While the experiments conducted in this study have all been performed well in simulation, in future work we plan to conduct more experiments with real robots to further evaluate and refine the proposed methods.

## 6 Conclusion and Discussion

We introduce Vlaser, a foundational vision-language-action model that extends vision-language models with embodied reasoning and end-to-end robot control capabilities. Powered by the Vlaser-6M dataset, the model establishes a new state of the art across a wide range of embodied reasoning benchmarks, including planning, grounding, spatial reasoning, and simulation-based tasks. Moreover, Vlaser reveals the most effective data streams for downstream VLA through its curated data pipeline, achieving state-of-the-art performance on Bridge and competitive results on Google Robot for end-to-end robot control.

In this work, we reveal that current embodied reasoning benchmarks exhibit a significant domain gap when compared to real-world robots. This core domain shift arises from the observation that robots have a fundamentally different viewpoint from that of internet datasets. Additionally, there are inherent limitations due to the lack of sufficient data from the robot's perspective, despite the abundance of vision datasets available. Therefore, we argue that it is essential to develop alignment techniques to bridge the domain gap in representations between the robot's viewpoint and that of internet datasets.

## ACKNOWLEDGMENTS

This work is supported by Shanghai Artificial Intelligence Laboratory.

## ETHICS STATEMENT

This research adheres to the ICLR 2026 ethical guidelines and upholds the principles of responsible research. We ensure that no personally identifiable, sensitive, or harmful data were used. Our experiments were based on publicly available datasets and did not involve any human subjects or vulnerable groups. We have considered the potential societal impact of our methods, including the risk of misuse, and believe that these contributions primarily advance scientific understanding and do not pose foreseeable harm.

## REPRODUCIBILITY STATEMENT

We follow the reproducibility guidelines in the ICLR 2026 author guidelines. We will open source code, configuration files, and scripts to reproduce our results, including dataset construction, model training, and evaluation, on platforms such as GitHub and Huggingface as soon as possible.

## REFERENCES

Sonnet Anthropic. Claude 3.7 sonnet system card. 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $pi\_\{0\}$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023a. URL https://arxiv.org/abs/2307.15818.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*, 2023b.

Junting Chen, Haotian Liang, Lingxiao Du, Weiyun Wang, Mengkang Hu, Yao Mu, Wenhai Wang, Jifeng Dai, Ping Luo, Wenqi Shao, et al. Owmm-agent: Open world mobile manipulation with multi-modal agentic data synthesis. *arXiv preprint arXiv:2506.04217*, 2025a.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025b.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025c.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *CoRR*, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

Long Cheng, Jiafei Duan, Yi Ru Wang, Haoquan Fang, Boyang Li, Yushan Huang, Elvis Wang, Ainaz Eftekhar, Jason Lee, Wentao Yuan, et al. Pointarena: Probing multimodal grounding through language-guided pointing. *arXiv preprint arXiv:2505.09990*, 2025.

Ron Chrisley. Embodied artificial intelligence. *Artificial intelligence*, 149(1):131–150, 2003.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025.

Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Huang Fang, Mengxi Zhang, Heng Dong, Wei Li, Zixuan Wang, Qifeng Zhang, Xueyun Tian, Yucheng Hu, and Hang Li. Robix: A unified model for robot interaction, reasoning and planning. *arXiv preprint arXiv:2509.01106*, 2025.

Alessandro Gasparetto, Paolo Boscariol, Albano Lanzutti, and Renato Vidoni. Path planning and trajectory planning algorithms: A general overview. *Motion and operation planning of robotic systems: Background and practical approaches*, pp. 3–27, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Zhi Hou, Tianyi Zhang, Yuwen Xiong, Haonan Duan, Hengjun Pu, Ronglei Tong, Chengyang Zhao, Xizhou Zhu, Yu Qiao, Jifeng Dai, et al. Dita: Scaling diffusion transformer for generalist vision-language-action policy. *arXiv preprint arXiv:2503.19757*, 2025.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pp. 9118–9147. PMLR, 2022.

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.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 1724–1734, June 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22225–22233, 2025.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Kanxue Li, Baosheng Yu, Qi Zheng, Yibing Zhan, Yuhui Zhang, Tianle Zhang, Yijun Yang, Yue Chen, Lei Sun, Qiong Cao, Li Shen, Lusong Li, Dapeng Tao, and Xiaodong He. Muep: A multimodal benchmark for embodied planning with foundation models. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 129–138. International Joint Conferences on Artificial Intelligence Organization, 8 2024a. doi: 10.24963/ijcai.2024/15. URL https://doi.org/10.24963/ijcai.2024/15. Main Track.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024b.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024c.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024d.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023a. URL https://arxiv.org/abs/2210.02747.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Huaping Liu, Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, and Hanbo Zhang. Towards generalist robot policies: What matters in building vision-language-action models. 2025.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=IL5zJqfxAa.

Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). In *European Conference on Computer Vision*, pp. 264–273. Springer, 2024.

NVIDIA, Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, Jingyi Jin, Brendan Johnson, Rizwan Khan, George Kurian, Elena Lantz, Nayeon Lee, Zhaoshuo Li, Xuan Li, Tsung-Yi Lin, Yen-Chen Lin, Ming-Yu Liu, Andrew Mathau, Yun Ni, Lindsey Pavao, Wei Ping, David W. Romero, Misha Smelyanskiy, Shuran Song, Lyne Tchapmi, Andrew Z. Wang, Boxin Wang, Haoxiang Wang, Fangyin Wei, Jiashu Xu, Yao Xu, Xiaodong Yang, Zhuolin Yang, Xiaohui Zeng, and Zhe Zhang. Cosmos-reason1: From physical common sense to embodied reasoning, 2025a. URL https://arxiv.org/abs/2503.15558.

NVIDIA, Nikita Cherniadev Johan Bjorck andFernando Castañeda, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. In *ArXiv Preprint*, March 2025b.

OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2025.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.

Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, et al. Embodiedonevision: Interleaved vision-text-action pretraining for general robot control. *arXiv preprint arXiv:2508.21112*, 2025a.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025b.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7141–7151, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017a. URL https://openreview.net/forum?id=B1ckMDqlg.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017b.

Junhao Shi, Zhaoye Fei, Siyin Wang, Qipeng Guo, Jingjing Gong, and Xipeng Qiu. World-aware planning narratives enhance large vision-language model planner, 2025. URL https://arxiv.org/abs/2506.21230.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020.

Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U Tan, Deepanway Ghosal, Soujanya Poria, et al. Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning. *arXiv preprint arXiv:2412.11974*, 2024.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 251–266. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/021bbc7ee20b71134d53e20206bd6feb-Paper.pdf.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=u6imHU4Ebu.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025a.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025b.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025c.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023a.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *arXiv preprint arXiv:2305.03716*, 2023.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14203–14214, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025c.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025d.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *arXiv preprint arXiv:2508.13998*, 2025.

Han-ye Zhang, Wei-ming Lin, and Ai-xia Chen. Path planning for the mobile robot: A review. *Symmetry*, 10(10):450, 2018.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025.

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. *arXiv preprint arXiv:2412.18194*, 2024.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025a.

Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Ran Cheng, Yaxin Peng, Chaomin Shen, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. *arXiv preprint arXiv:2502.14420*, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Allen Zren. open-pi-zero: Re-implementation of pi0 vision-language-action (vla) model, 2025. URL `https://github.com/allenzren/open-pi-zero`.

# A Appendix

## THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used a large language model, ChatGPT (OpenAI, 2025), only for grammar check and correction. All the writings was manually reviewed by the authors. Crucially, AI was not used to generate any research data, statistical analyses, figures, or conclusions. The authors take full responsibility for the final content and any errors herein.

## A.1 Training Details

Our Vlaser is optimized in a fully supervised finetuning (SFT) manner based on InternVL3 series (Zhu et al., 2025)(InternVL3-2B and InternVL3-8B). To maximize adaptation to embodied reasoning tasks, we keep all parameters trainable, including those in the large language model, the vision-language projector, and the visual encoder, enabling comprehensive end-to-end learning. Further details regarding the training setup, including hyperparameters and optimization settings, are provided in Table 6.

Table 6: **Hyper-parameters used in the VLM pretraining of Vlaser.**

| Configurations | Values |
|---|---|
| LLM sequence length | 16, 384 |
| Dynamic Resolution | True |
| Patch Size | 448 |
| Max Patch num | 12 |
| Freeze vision tower | False |
| Freeze multimodal projector | False |
| Freeze language model | False |
| Optimizer | AdamW |
| Optimizer hyperparameters | $\beta_1 = 0.9, \beta_2 = 0.999, eps = 1e-8$ |
| Peak learning rate | 2e-5 |
| Learning rate schedule | cosine decay |
| Training epochs | 1 |
| Training steps | 5, 000 |
| Warm-up steps | 150 |
| Global batch size | 128 |
| Gradient accumulation | 2 |
| Numerical precision | bfloat16 |

While using Vlaser as the base model for downstream VLA Policy fine-tuning, we optimize all parameters within both the VLM and the Action Expert. Additionally, we conduct comparative experiments using several different versions of base models, including InternVL3-2B, etc. Detailed information and related parameter settings can be found in Table 7.

## A.2 Data Generation Details

**Embodied Grounding Data** To further enhance embodied grounding capabilities, we generate an additional 300k high-quality data samples from the SA-1B dataset (Kirillov et al., 2023). The

Table 7: **Hyper-parameters used in the VLA fine-tuning.**

| Configurations | Values |
|---|---|
| LLM sequence length | 384 |
| Image Size | 448 |
| Freeze VLM | False |
| Global batch size | 1024 |
| Training epochs | 10 |
| VLM Peak Learning Rate | 5e-5 |
| Action Expert Peak Learning Rate | 5e-5 |
| Optimizer | AdamW |
| Optimizer hyperparameters | $\beta_1 = 0.9, \beta_2 = 0.999, eps = 1e-8$ |
| Observation history length | 1 |
| Action Chunk length | 4 |
| Execute Action length | 2 |

data generation process consists of two main stages. First, we convert segmentation masks into bounding boxes and point annotations: bounding boxes are derived by computing the minimal axis-aligned rectangle enclosing each mask, while point annotations are obtained by randomly sampling a coordinate within the mask region. To ensure annotation quality, we apply an IoU threshold of 0.9 to select high-precision masks; masks with lower IoU values are either excluded or assigned reduced sampling weight. From the over 1 billion available masks, we initially sample 1 million candidate instances. In the second stage, we employ a two-step captioning and refinement pipeline. Coarse captions are first generated using BLIP-2 (Li et al., 2023), followed by a filtering and refinement process using Qwen2.5-VL-7B (Bai et al., 2025) to eliminate low-quality items and produce more accurate and detailed descriptions. This rigorous pipeline ultimately yields 300k high-quality data samples tailored for embodied grounding tasks, significantly expanding the diversity and precision of our training corpus.

**Spatial Reasoning Data** To enhance spatial intelligence capabilities, we manually construct a dataset of 100k 3D spatial perception samples derived from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021). Following methodologies established in prior data engines (Deng et al., 2025; Fan et al., 2025), we utilize both the 3D point cloud and video sequences of each scene to construct a spatio-temporal scene graph. This graph encapsulates structural and semantic information such as overall scene dimensions, room center coordinates, object category counts, and precise 3D bounding boxes for every object instance. Based on this representation, we generate spatial reasoning questions that probe layout properties and inter-object relationships. These include queries about the object counts, absolute and relative distances, object and room sizes, relative directions between objects, and other spatial attributes, using the same question template in VSI-Bench (Yang et al., 2025b).

**Planning Data** To improve the model's ability to comprehend complex instructions and execute tasks with environmental feedback, we curate additional planning data within the Habitat simulator (Szot et al., 2021). Specifically, we initialize each planning task following the annotations of LLaRP (Szot et al., 2024), which specify both the task goals and the set of permissible actions. An LLM agent based on gpt-4o (OpenAI, 2025) is then deployed to roll out the task. During each rollout, we record the task instruction, the sequence of actions taken, and the environment's feedback, including observations and success signal for each action. Both the executed action trajectories and the corresponding feedback are retained. Only trajectories that successfully accomplish the task are included in the final training set, providing rich paired data of instructions, execution processes, and environment responses for enhancing the model's planning capabilities in a complex environment.

**In-Domain Data for downstream VLAs** We generate 2 million in-domain multimodal data samples to facilitate direct transfer learning for downstream Vision-Language Agents (VLAs) during fine-tuning. These data are collected from two distinct simulation platforms: the WidowX Robot (Walke et al., 2023a) and the Google Robot (Brohan et al., 2023b;a) within the SimplerEnv (Li et al., 2024d), as well as the Aloha-AgileX Robot from RoboTwin2.0 (Chen et al., 2025c). The dataset mainly encompasses three systematically designed question-answer types: 1) General QA, which queries the robot's current state and requests next few action plans; 2) Grounding QA, which requires the robot to localize points and bounding boxes as the actionable affordances; 3) Spatial Reasoning

QA, which probes relative spatial relationships and geometric properties of objects in the scene. Detailed prompt templates and representative examples for each data category are provided in Figure 3 (General QA), Figure 4 (Grounding QA), and Figure 5 (Spatial Reasoning QA), respectively. We use Qwen2.5VL-7B (Bai et al., 2025) as the base model to generate the above data items.

To further enhance the quality and diversity of the generated in-domain QA data, we implement a post-processing data filtering pipeline. In line with established practices in dataset construction (Wang et al., 2023; Cui et al., 2023; Li et al., 2024b), we employ an LLM-as-a-judge approach (Li et al., 2025) to score each generated data sample on a scale of 1 to 3, using Qwen2.5VL-32B (Bai et al., 2025) as the judge model. The detailed instruction prompts for General QA, Spatial QA, and Grounding QA are provided below. We filter out all samples assigned a score of 1, which account for approximately 10% of the initially generated data. This outcome reflects the overall high quality and diversity of our in-domain dataset. To further validate the reliability of the LLM-based evaluation, we randomly selected a subset of the scored data (covering all score levels from 1 to 3) for human reassessment by three human experts. The results show that the LLM judge's scores align with human ratings in nearly 80% of cases, confirming the consistency and reliability of our automated evaluation and data filtering pipeline.

### Prompt of Data filtering for General QA data

**You are an expert in evaluating question-answer pairs for robot arm camera images and task instructions. Please evaluate the quality of the generated QA pair based on Relevance, Informativeness, and Clarity.**

You will receive the following inputs: — The robot arm camera image, the task instruction, the generated question, and the generated answer. Carefully evaluate whether the question and answer are relevant to the image and task, informative, and clearly expressed.

**Scoring Criteria (1–3):**

- **1 – Poor** The QA pair is irrelevant, uninformative, or unclear. - The question or answer doesn't relate to the image or task instruction. - The answer is vague, generic, or provides no useful information. - The question is poorly formulated or doesn't make sense.

  **Examples:** - Question: "What is this?" Answer: "It's an image." (too generic) - Question: "How to cook?" Answer: "Use a pan." (irrelevant to robot task) - Question: "What color?" Answer: "Color." (unclear and uninformative)
- **2 – Fair** The QA pair is somewhat relevant but lacks depth or clarity. - The question relates to the image/task but is too simple or generic. - The answer provides basic information but lacks detail or specificity. - Some aspects are unclear or could be better explained.

  **Examples:** - Question: "What objects are in the image?" Answer: "There are some objects." (too vague) - Question: "What is the robot doing?" Answer: "The robot is moving." (lacks detail) - Question: "How to complete the task?" Answer: "Follow the instructions." (not specific enough)
- **3 – Good** The QA pair is highly relevant, informative, and clearly expressed. - The question is specific, well-formulated, and directly relates to the image and task. - The answer provides detailed, accurate, and useful information. - Both question and answer are clear and help understand the robot's environment and task.

  **Examples:** - Question: "What objects are visible in the robot arm's workspace and which one should be manipulated based on the task instruction?" Answer: "I can see a red cup, a blue bowl, and a green plate on the table. According to the task instruction to 'pick up the cup', the robot should focus on the red cup located in the center of the workspace." - Question: "What obstacles might prevent the robot from completing the task?" Answer: "The workspace appears clear, but the target object is positioned near the edge of the table, which may require careful positioning to avoid knocking it over during manipulation."

**Output Requirement:** You must return only a single integer score from 1 to 3. Do not include any explanation, labels, or extra content.
**Question:** {question}
**Answer:** {answer}

### Prompt of Data filtering for Spatial QA data

**You are an expert in evaluating spatial intelligence question-answer pairs for robot arm camera images and task instructions. Please evaluate the quality based on Spatial Reasoning Accuracy, Detail Level, and Relevance.**

You will receive the following inputs: — The robot arm camera image, the task instruction, the generated spatial question, and the generated spatial answer.
Carefully evaluate whether the question targets spatial reasoning, and whether the answer provides accurate and detailed spatial information.

**Scoring Criteria (1–3):**

- **1 – Poor** The spatial QA pair lacks spatial reasoning focus or provides incorrect/vague spatial information. - The question doesn't target spatial aspects (counting, relationships, distances, orientation, etc.). - The answer provides incorrect spatial information or is too vague. - The spatial reasoning is flawed or irrelevant to the task.

**Examples:** - Question: "What is in the image?" Answer: "Objects." (not spatial) - Question: "How many objects?" Answer: "Some." (vague, no specific count) - Question: "Where is object A?" Answer: "It's there." (no spatial detail) - Question: "What is the distance?" Answer: "Close." (not quantitative)

- **2 – Fair** The spatial QA pair addresses spatial reasoning but lacks precision or detail. - The question targets spatial aspects but could be more specific. - The answer provides basic spatial information but lacks quantitative details or precision. - Some spatial relationships are mentioned but not fully explained.

  **Examples:** - Question: "How many cups are there?" Answer: "There are a few cups." (not specific count) - Question: "Where is the red object relative to the blue one?" Answer: "The red one is near the blue one." (lacks specific direction/distance) - Question: "What is the spatial arrangement?" Answer: "Objects are arranged on the table." (too general)

- **3 – Good** The spatial QA pair demonstrates strong spatial reasoning with precise, detailed information. - The question clearly targets specific spatial aspects (counting, relationships, distances, orientation, geometry, etc.). - The answer provides accurate, quantitative, and detailed spatial information. - The spatial reasoning is relevant to the robot task and helps understand the workspace layout.

  **Examples:** - Question: "How many objects are visible in the scene and what are their types?" Answer: "I can count 5 objects: 2 red cups positioned on the left side of the table, 1 blue bowl in the center, and 2 green plates arranged on the right side." - Question: "What is the relative position of the target object compared to the robot arm's current position?" Answer: "The target object (red cup) is located approximately 30cm to the right and 15cm forward from the robot arm's current end-effector position, requiring a diagonal reach motion." - Question: "Which objects are within the robot's reachable workspace?" Answer: "Based on the robot arm's reach, the blue bowl (center, 25cm away) and the left red cup (20cm away) are within reach. The rightmost green plate (45cm away) is outside the immediate reachable zone."

**Output Requirement:** You must return only a single integer score from 1 to 3. Do not include any explanation, labels, or extra content.
**Question:** {question}
**Answer:** {answer}

## Prompt of Data filtering for Grounding QA data

**You are an expert in evaluating visual grounding question-answer pairs for robot arm camera images and task instructions. Please evaluate the quality based on Grounding Accuracy, Coordinate Validity, and Localization Precision.**

You will receive the following inputs: — The robot arm camera image, the task instruction, the generated grounding question, and the generated grounding answer (which should contain coordinates or bounding boxes). Carefully evaluate whether the question targets object localization, and whether the answer provides valid and accurate coordinate information.

**Scoring Criteria (1–3):**

- **1 – Poor** The grounding QA pair lacks localization focus or provides invalid/incorrect coordinate information. - The question doesn't target object localization, pointing, or detection. - The answer lacks coordinate information or contains invalid coordinates (out of bounds, wrong format). - The coordinates don't match the described object location.

  **Examples:** - Question: "What is in the image?" Answer: "A cup." (no coordinates) - Question: "Where is the object?" Answer: "It's on the table." (no coordinates) - Question: "Point to the cup." Answer: "The cup is at (1500, 2000)." (coordinates out of bounds for normalized 0-1000 range) - Question: "Find the object." Answer: "Box: [100, 200, 50, 300]." (invalid box format, x2 < x1)

- **2 – Fair** The grounding QA pair addresses localization but coordinates are imprecise or partially valid. - The question targets localization but could be more specific. - The answer contains coordinates but they may be approximate, slightly off, or lack precision. - Some coordinate information is present but formatting could be improved.

  **Examples:** - Question: "Where is the object?" Answer: "The object is at position (500, 300)." (single point, but should specify if multiple objects exist) - Question: "Point to all cups." Answer: "Cups are at (200, 150) and (600, 400)." (coordinates present but may not be precise) - Question: "Mark the boundaries." Answer: "Box: [100, 100, 400, 300]." (valid box but may not accurately bound the object)

- **3 – Good** The grounding QA pair demonstrates precise localization with valid and accurate coordinate information. - The question clearly targets object localization, pointing, detection, or precise positioning. - The answer provides valid, accurate coordinates (points or bounding boxes) in the correct format. - Coordinates are normalized to 0-1000 range and accurately represent object locations. - Multiple objects are properly localized if applicable.

  **Examples:** - Question: "Where is the red cup located in the image? Provide coordinates." Answer: "The red cup is located at point (450, 320). <point>[[450, 320]]</point>" - Question: "Point to all instances of cups visible in the scene." Answer: "I can locate 2 cups: the red cup at (450, 320) and the blue cup at (680, 250). <point>[[450, 320], [680, 250]]</point>" - Question: "Can you locate and mark the boundaries of the target object?" Answer: "The target object (red cup) is bounded by box [380, 280, 520, 360]. <box>[[380, 280, 520, 360]]</box>" - Question: "Find and mark all instances of plates in the image." Answer: "I found 2 plates: plate 1 at box [100, 200, 250, 350], plate 2 at box [600, 180, 750, 330]. <box>[[100, 200, 250, 350], [600, 180, 750, 330]]</box>"

**Output Requirement:** You must return only a single integer score from 1 to 3. Do not include any explanation, labels, or extra content.
**Question:** {question}
**Answer:** {answer}

### A.3 SIMULATION EVALUATION DETAILS

We fine-tune and evaluate the VLA models using various base models within the SimplerEnv simulation environment. To ensure fair evaluation, we use checkpoints with the same number of iterations for the WidowX Robot Task and the Google Robot Task, respectively. Specifically, for the WidowX Robot Tasks, we use a checkpoint after 45,390 iterations, while for the Google Robot Tasks, we use a checkpoint after 36,970 iterations. During evaluation, we utilize a single image and select an action chunk of size 2 for execution. In the flow matching configuration, we use 10 inference steps during the inference phase and apply Euler method as numerical integration method. We evaluate a sufficient number of samples to ensure the reliability of the tests. The exact number of test samples for each task is shown in the Table 8.

Table 8: **The number of samples evaluated on SimplerEnv.**

| | | Task Name | Evaluation Samples |
|---|---|---|---|
| Widowx Robot Tasks | | Carrot on plate | 240 |
| | | Put eggplant in basket | 240 |
| | | spoon on towel | 240 |
| | | stack cube | 240 |
| Google Robot Tasks | Visual Matching | Pick coke can | 300 |
| | | Move Near | 240 |
| | | Open/Close Drawer | 216 |
| | Variance Aggregation | Pick coke can | 825 |
| | | Move Near | 600 |
| | | Open/Close Drawer | 378 |

### A.4 QUALITATIVE DEMONSTRATION

**Qualitative Samples for Embodied Reasoning**. To gain in-depth insights into Vlaser, we provide extensive visualizations on embodied reasoning tasks including Embodied QA 9, grounding 10, spatial intelligence 11 and planning tasks in Simulation 12. These examples not only validate the generalizability of Vlaser, but also confirm the superior capabilities of Vlaser across all tasks.

**Qualitative Samples in SimplerEnv**. Figure 6 and Figure 7 present qualitative examples from the evaluations conducted in the SimplerEnv. Specifically, Figure 6 showcases the performance of multiple models on the WidowX Robot Task, while Figure 7 demonstrates the results on the Google Robot Task. From these results, it is evident that, under the same test conditions, baseline models often fail the entire task due to slight positional errors. In contrast, our model shows significant improvements in this regard. Whether incorporating in-domain question-answer pair data, spatial intelligence data, or grounding data, the performance of the model shows clear enhancement. These results validate the effectiveness of our approach.

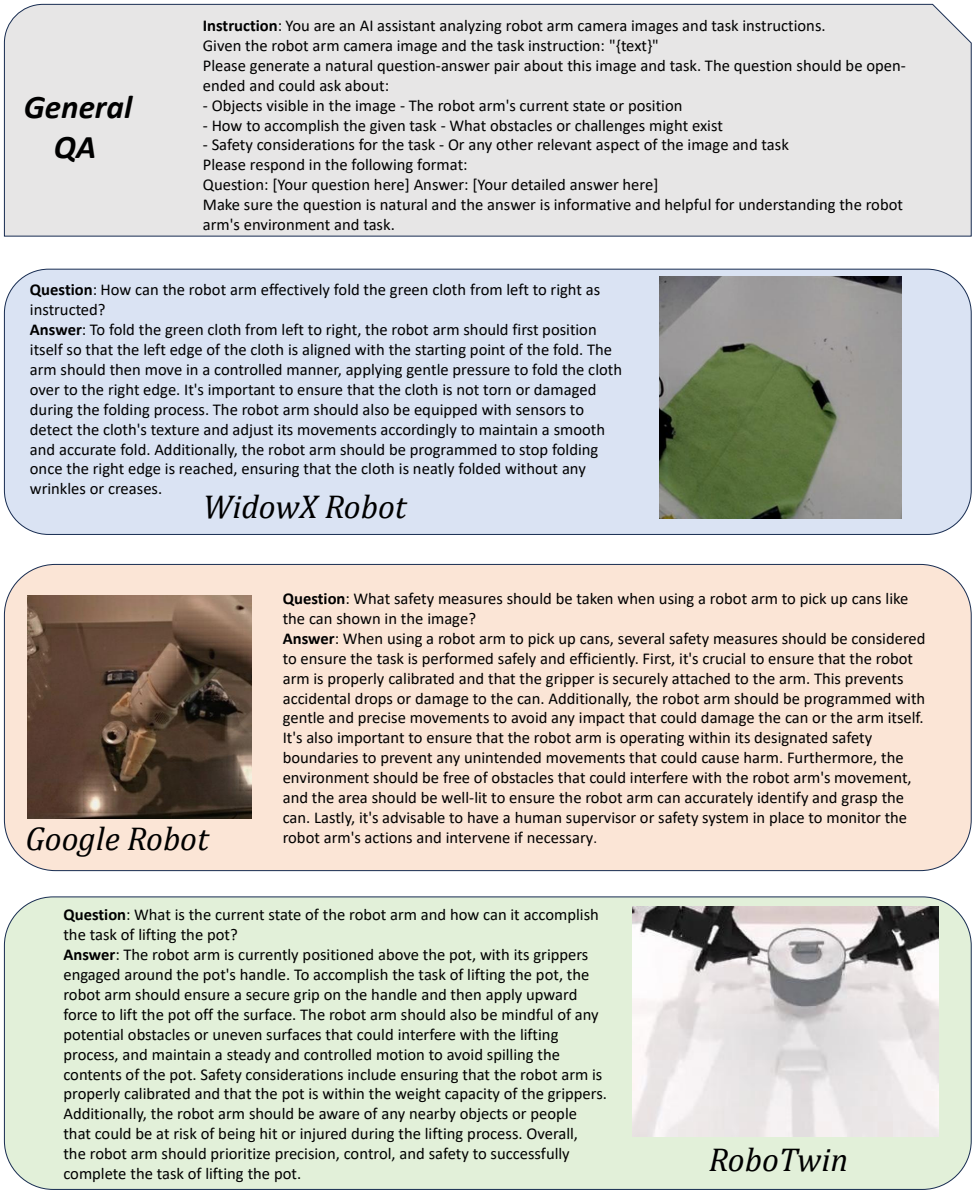



Figure 3: An illustration of Vlaser-6M data engine for in-domain general QA sample in SimplerEnv and RoboTwin.

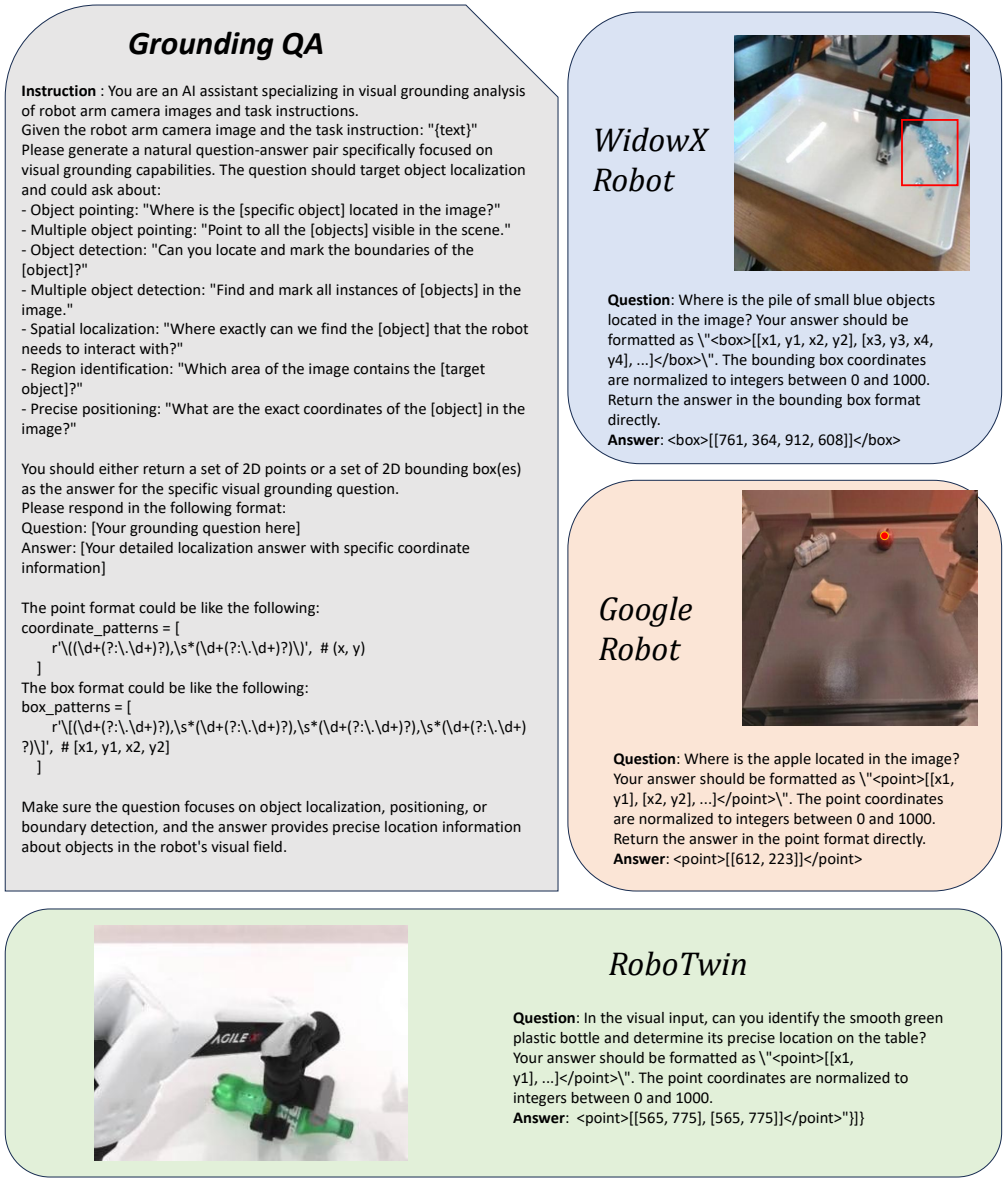



Figure 4: An illustration of Vlaser-6M data engine for in-domain embodied grounding QA sample in SimplerEnv and RoboTwin.

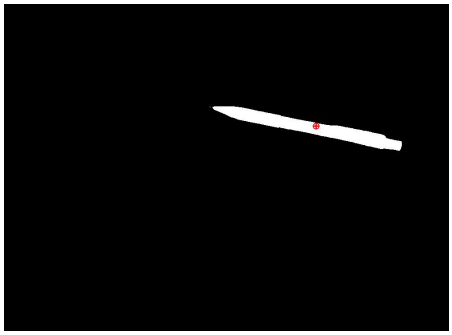
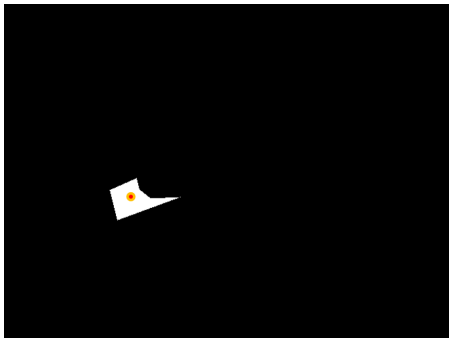

Figure 5: An illustration of Vlaser-6M data engine for in-domain spatial reasoning QA sample in SimplerEnv and RoboTwin.

Table 9: Embodied QA examples.

| **Embodied QA: Example #1 from ERQA.** | |
|---|---|
|   | |
| **Question**: | What happened between these two frames? Choices: A. Robot arm lifted the cup. B. Robot arm poured all the nuts into a cup. C. Robot arm poured some of the nuts into a cup. D. Nothing happened. Please answer directly with only the letter of the correct option and nothing else. |
| **Vlaser-8B**: | C. |

| **Embodied QA: Example #2 from ERQA.** | |
|---|---|
|     | |
| **Question**: | Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option. Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to weigh dough? A. put dough on table B. pick dough C. roll dough on table D. cut dough with dough cutter |
| **Vlaser-8B**: | A. |

Table 10: Embodied Grounding Examples.

| **Embodied Grounding: Example #1 from PointArena.** | |
|---|---|
|  |  |
| **Question**: | Point to the tool people use for writing. |
| **Vlaser-8B**: | <point>[[701, 374]]</point>. |
| **Embodied Grounding: Example #2 from Where2Place.** | |
|  |  |
| **Question**: | Please point out the free space between the black water bottle and the pot lid. |
| **Vlaser-8B**: | <point>[[293, 560]]</point>. |

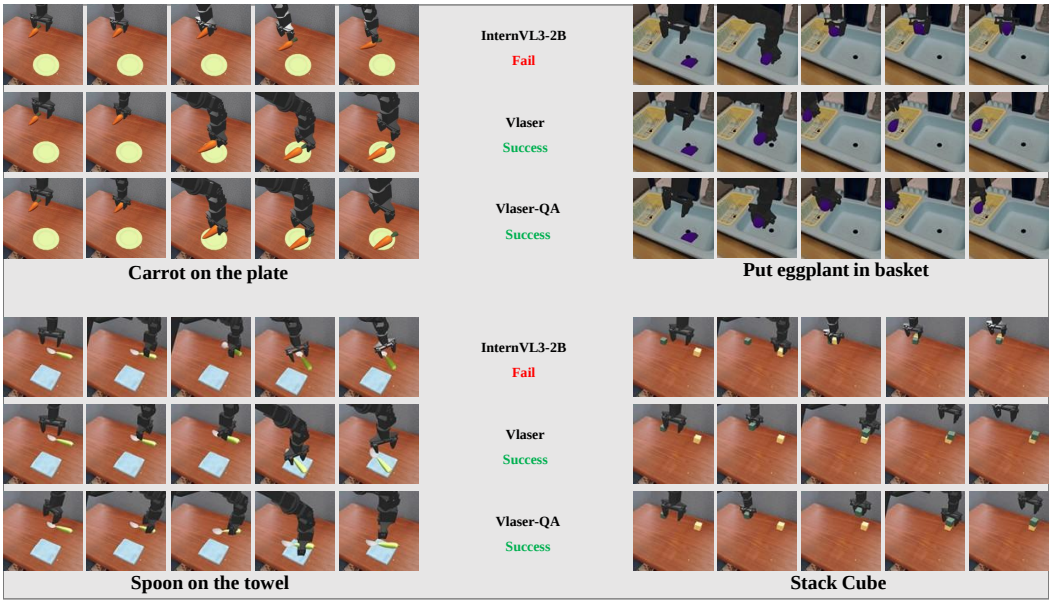



Figure 6: Qualitative samples in SimplerEnv on WidowX Robot Tasks

Table 11: Spatial Intelligence Examples.

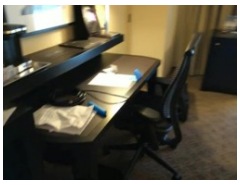
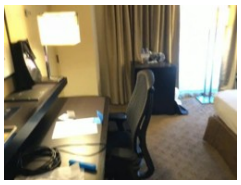
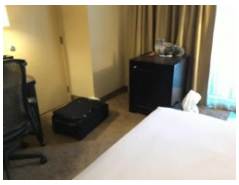
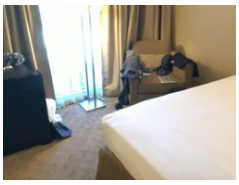
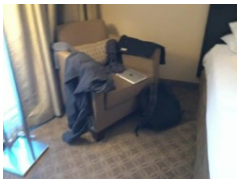
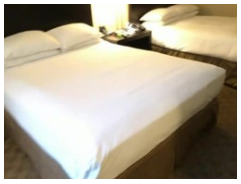
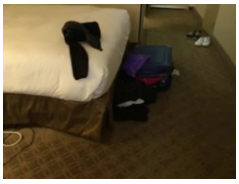
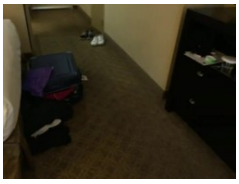
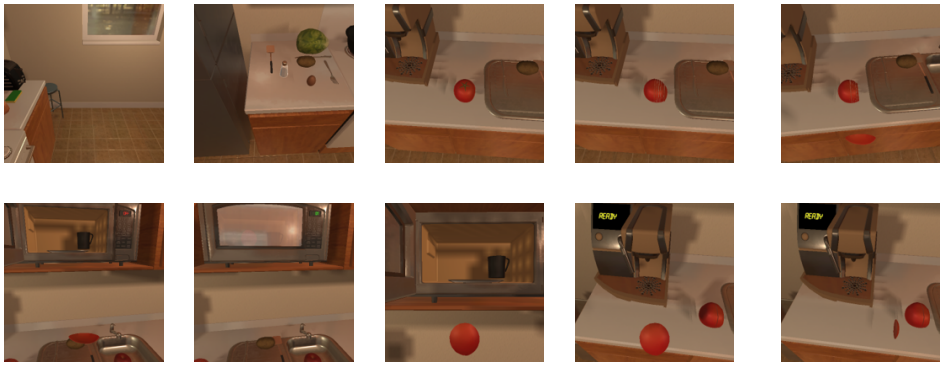
| **Spatial Reasoning: Example #1 from VSI.** | |
|---|---|
|         | |
| **Question**: | You are a robot beginning at the standing by the window and facing the window. You want to navigate to the white shoes. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward passing the bed 3. [please fill in] 4. Go forward until the white shoes. You have reached the final destination. A. Turn Right, Turn Left B. Turn Left, Turn Left C. Turn Left, Turn Right D. Turn Right, Turn Right. Answer with the option's letter from the given choices directly. |
| **Vlaser-8B**: | B. |

Table 12: Simulation planning Examples.

| **Simulation Planning: Example #1 from EmbodiedBench.** | |
|---|---|
|  | |
| **Question**: | You are a robot operating in a home. Given a task, you must accomplish the task using a defined set of actions to achieve the desired outcome. Now the human instruction is: Put a microwaved tomato slice on the counter. You need to output the reasoning process and an action to execute. |
| **Vlaser-8B**: | Reasoning-step-1: To put a microwaved tomato slice on the counter, I need to start by slicing a tomato. To do this, I require a knife. In the image, there is a knife visible on the counter near the toaster. Finding the knife is the logical next step because it is essential for cutting the tomato into slices. Once I have the knife, I can proceed with slicing the tomato. Therefore, my next action is to find the knife. Action: find a knife.<br>Reasoning-step-2: I have identified the knife on the counter near the sink. Picking up the knife is the next logical step because it is necessary for slicing the tomato. Once I have the knife, I can proceed with cutting the tomato into slices. This aligns with the goal of preparing a tomato slice for microwaving and then placing it on the counter. Action: pick up the knife.<br>Reasoning-step-3: I have successfully located the knife in the drawer. Now, I need to find a tomato to slice. The image shows a kitchen counter with various items, but no tomato is immediately visible. Tomatoes are commonly stored in the refrigerator or on the counter. Since the refrigerator is partially visible on the left side of the image, it makes sense to check there first. If the tomato isn̈t in the fridge, I can then look around the counter more thoroughly. Therefore, my next step is to find a tomato. Action: find a tomato.<br>...<br>Reasoning-step-9: The current visual state shows the microwave with its door closed, and the tomato slice is on the counter. The task is to place a microwaved tomato slice on the counter, which has already been accomplished. Given that the tomato slice is already on the counter, the next logical step is to confirm its placement by finding a suitable counter top. This ensures that the tomato slice is securely and appropriately positioned for the task. Therefore, the next step is to find a counter top. Action: find a counter top.<br>Reasoning-step-10: The current visual state shows a microwave with its door closed, and two tomatoes are visible on the counter. One tomato is already sliced, indicating that the previous steps involving slicing and microwaving have been completed. Since the tomato is currently being held, the next logical step is to put it down on the counter. This action will free up the hand to handle the tomato slice and place it on the counter as required by the task. Therefore, the appropriate next action is to put down the tomato. Action: put down the tomato. |

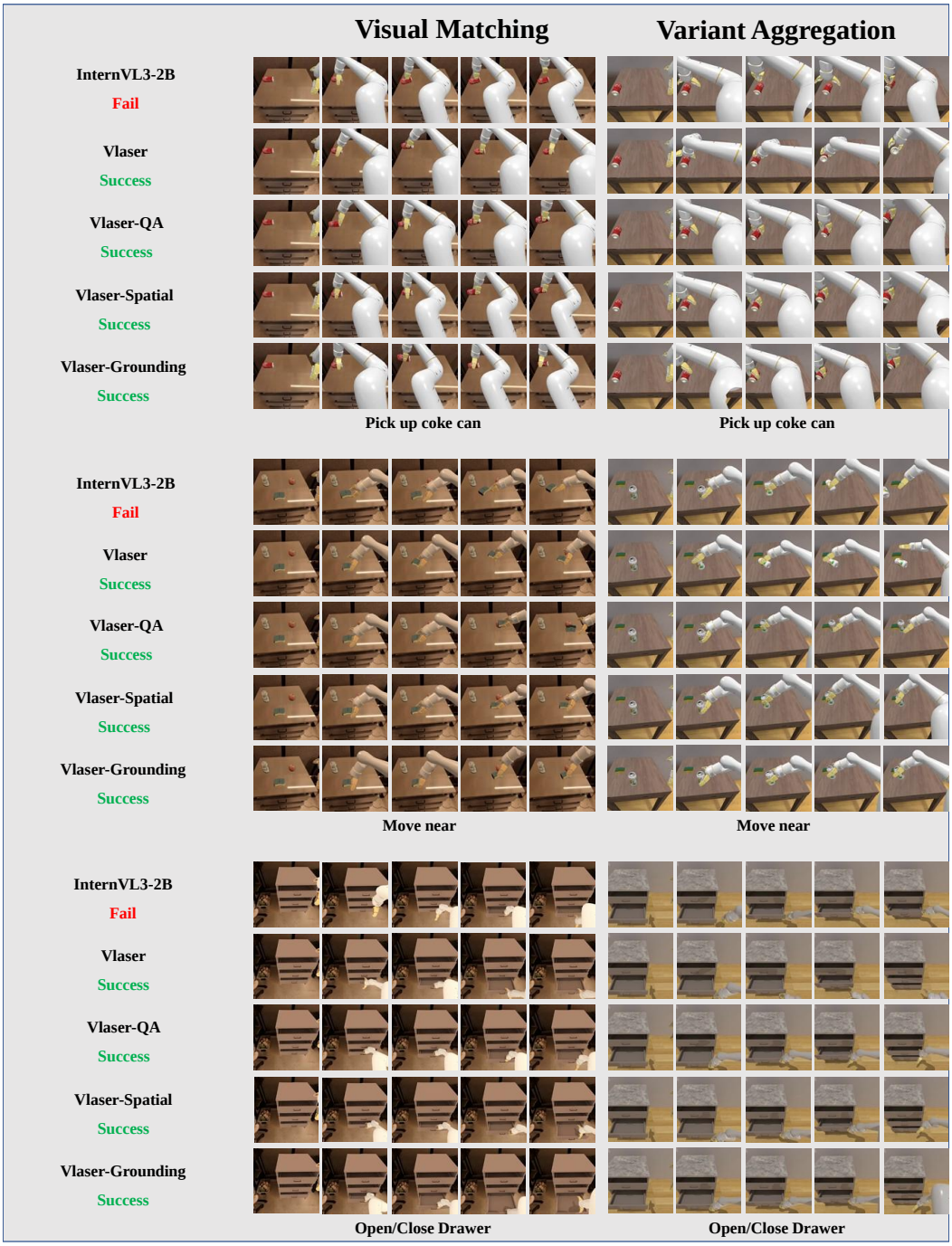



Figure 7: Qualitative samples in SimplerEnv on Google Robot Tasks