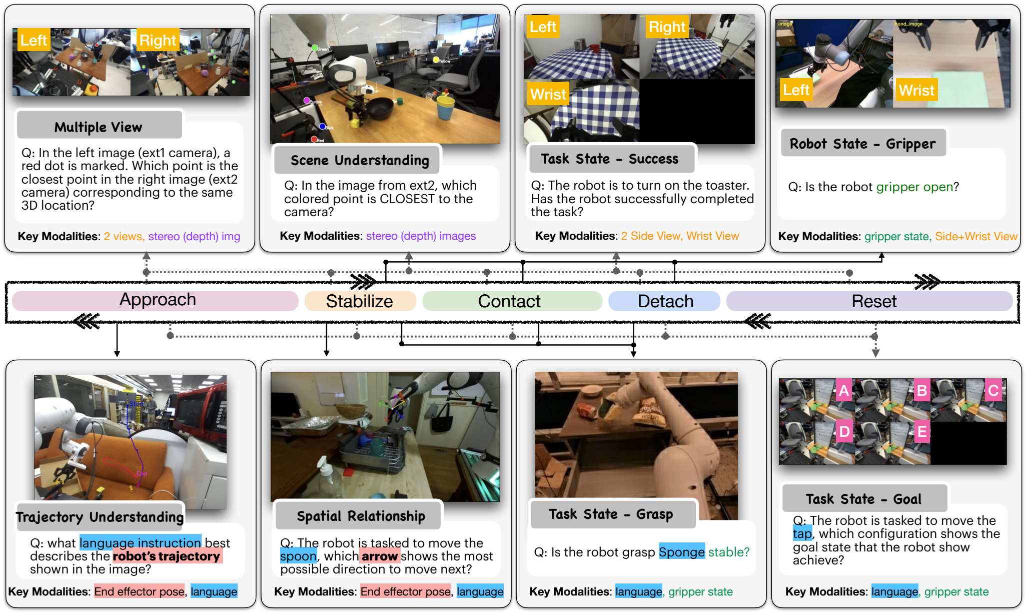
Robo2VLM:
Visual Question Answering from Large-Scale In-the-Wild Robot Manipulation Datasets

Kaiyuan Chen[1,*], Shuangyu Xie[1,*], Zehan Ma[1], Pannag R Sanketi[2], Ken Goldberg[1]
[1]University of California, Berkeley, [2]Google DeepMind, [*]Equal contribution
https://huggingface.co/datasets/keplerccc/Robo2VLM-1



Q: In the left image (ext1 camera), a red dot is marked. Which point is the closest point in the right image (ext2 camera) corresponding to the same 3D location?
Key Modalities: 2 views, stereo (depth) img

Q: In the image from ext2, which colored point is CLOSEST to the camera?
Key Modalities: stereo (depth) images

Q: The robot is to turn on the toaster. Has the robot successfully completed the task?
Key Modalities: 2 Side View, Wrist View

Q: Is the robot gripper open?
Key Modalities: gripper state, Side+Wrist View

Q: what language instruction best describes the **robot's trajectory** shown in the image?
Key Modalities: End effector pose, language

Q: The robot is tasked to move the spoon, which **arrow** shows the most possible direction to move next?
Key Modalities: End effector pose, language

Q: Is the robot grasp Sponge stable?
Key Modalities: language, gripper state

Q: The robot is tasked to move the tap, which configuration shows the goal state that the robot show achieve?
Key Modalities: language, gripper state

Fig. 1: **Robo2VLM-1 dataset overview**. The middle colorbar traces a typical manipulation episode—from pre-grasp through immobilization, contact, detach, and into post-grasp. Surrounding panels give example questions for each VQA category. Dashed arrows connect every category to the phase(s) in which its questions are sampled. Icons beneath each panel list the key sensing modalities (RGB, stereo depth, wrist/side cameras, gripper state, end-effector pose, language instructions) needed to derive ground-truth answers.

***Abstract*—Vision-Language Models (VLMs) acquire real-world knowledge and general reasoning ability through Internet-scale image-text corpora. They can augment robotic systems with scene understanding and task planning, and assist visuomotor policies that are trained on robot trajectory data. We explore the reverse paradigm — using rich, real, multi-modal robot trajectory data to enhance and evaluate VLMs. In this paper, we present Robo2VLM, a Visual Question Answering (VQA) dataset generation framework for VLMs. Given a human tele-operated robot trajectory, Robo2VLM derives ground-truth from non-visual and non-descriptive sensory modalities, such as end-effector pose, gripper aperture, and force sensing. Based on these modalities, it segments the robot trajectory into a sequence of manipulation phases. At each phase, Robo2VLM uses scene and interaction understanding to identify 3D properties of the robot, task goal, and the target object. The properties are used to generate representative VQA queries – images with textural multiple-choice questions – based on spatial, goal-conditioned, and interaction reasoning question templates. We curate Robo2VLM-1, a large-scale in-the-wild dataset with 684,710 questions covering 463 distinct scenes and 3,396 robotic manipulation tasks from 176k real robot trajectories. Results suggest that Robo2VLM-1 can benchmark and improve VLM capabilities in spatial and interaction reasoning.**

I. Introduction

Emerging Vision-Language Models (VLMs) [1–6] can perform high-level reasoning and scene interpretation [7, 8]. Recent robotic manipulation systems that integrate VLMs demonstrate enhanced capabilities in semantic and long horizon task reasoning [9–11]. Yet, *the* key challenge persists: the image-text corpora used for VLM pre-training high-quality lack fine-grained spatial information, which are prerequisites for robots to identify long-tail objects, complex scenes, reason about spatial relationships, and plan physical interactions.

To address this challenge, some research [12–14] relies on data generation through simulation [15–17]. However, such data has inherent limitations due to the sim-to-real gap, because simulator cannot accurately model visual properties such as noise, clutter, and lighting variations and physical properties such as contact dynamics, and interactions. Therefore, strong performance in simulation often fails to translate reliably to the physical world. Meanwhile, deriving spatial knowledge from real-world ("in-the-wild") data typically requires extensive and costly human labeling [18, 19]. In contrast, teleoperated robot trajectories that are used to train visuomotor policies [20], such as Vision-Language-Action(VLA) [9, 21] or diffusion policies [22], typically include precise, structured proprioceptive and kinematic information—joint angles, end-effector poses, gripper states, and force–torque readings—that implicitly encode 3D spatial information. We hypothesize that visual and textual data extracted from robot trajectories can improve VLM's spatial reasoning capabilities.

We present Robo2VLM, a multiple-choice Visual Question Answering (VQA) dataset generation framework for VLMs from real-world robot data. Given a human-teleoperated robot trajectory, Robo2VLM segments the trajectory into distinct manipulation phases, selects representative frames from each phase, and generates questions whose answers are supported by synchronized proprioceptive and kinematic ground truth. We apply Robo2VLM to 176k diverse, real-world trajectories from the Open X-Embodiment (OXE) dataset [23], producing over 3 million VQA samples. Inspired by data optimization paradigms such as domain reweighting in natural language processing [24] and robot policy learning [25], we curate Robo2VLM-1, a large-scale, in-the-wild VQA dataset with 684,710 questions covering 463 distinct scenes, 3,396 robotic manipulation tasks, and 149 manipulation skills.

We evaluate 14 model configurations with state-of-the-art open source models (LLaVA, Llama and Qwen) and with different parameter sizes and prompting techniques. The results indicate that some VLMs can achieve near human performance in questions related to object reachability and interaction understanding. Evaluation also suggests a significant gap to human performance, especially in complex reasoning of fine-grained spatial relationship and interactions. Finetuning LLaVA [3] with Robo2VLM-1 improves most of the spatial and interaction capabilities with increasing training dataset size, with a maximum 50% accuracy gain in state reasoning and task understanding.

This paper makes the following contributions: (1) Robo2VLM, a VQA data generation framework from real robot trajectories. (2) Robo2VLM-1, an open VQA dataset with 684,710 questions covering diverse and realistic evaluation scenarios for manipulation. (3) Extensive evaluation data on state-of-the-art and fine-tuned VLMs.

II. Related Work

Large-Scale Robotics Datasets: Recent large-scale robotics datasets, such as Open-X-Embodiment [23] and DROID [26], provide extensive teleoperated demonstrations of complex manipulation skills. These datasets are foundational for training modern generalist robot policies—including Octo [21], RT-1 [27], RT-2 [28], OpenVLA [9], Gemini Robotics [10], π_0 [29], and Hi Robot [11]—enabling them to learn diverse skills and understand nuanced physical interactions from broad data. Crucially for grounding VLMs, robotics datasets from Open-X-Embodiment contains rich sensory-modal including RGB video, proprioceptive [27, 30–43], depth data [27, 30–32, 34], and force-torque [36, 38–40], that reflect the dynamics of interaction. These information presents an opportunity to bridge robotics data with VLMs.

VQA Benchmarks for Robotics and Embodied AI: VQA offers a powerful paradigm for evaluating the visual reasoning capabilities of VLMs [44–46]. Recently, VQA benchmarks have been developed for robotic tasks such as visual navigation in long-horizon planning [47, 48]. Simulation-based approaches [12–14] (often utilizing environments like [15–17]) generate large-scale VQA dataset, but face the persistent sim-to-real domain gap, where the result may not hold in reality due to factors like noise, clutter, and lighting variations. Real-world data benchmark, such as RoboVQA [18] (human-verified Q/A), improve generalization to real world setting but often involve significant manual annotation effort. These methods typically do not fully automate VQA generation by exploiting the rich spectrum of non-visual modalities (e.g., force, torque, proprioception), limiting their ability to support questions grounded in concepts such as grasp stability or multi-view spatial alignment. In contrast, Robo2VLM reduces the need for manual annotation and enables interaction and physical properties reasoning that are underexplored in previous VQA benchmarks, such as gripper states, grasping stability, task goal, and spatial information focus on the robot and target objects.

III. Robo2VLM

Robo2VLM generates five-way multiple-choice question answering (MCQ) from real robot teleoperated trajectories. Robo2VLM offers the following key features: (1) High-quality and representative keyframe selection from long-horizon, in-the-wild, multi-modal robot trajectories, ensuring semantic diversity and relevance; (2) Manipulation-centric question generation encompassing spatial, goal-conditioned, and interaction reasoning, each aligned with specific manipulation phases and grounded in corresponding sensor modalities.

We begin by defining a robot trajectory as a time-synchronized sequence of data frames from multiple sensor modalities including exteroceptive and proprioceptive [49]. Let T denote the length of a trajectory, and let $t \in \{1, 2, \ldots, T\}$ index the discrete time steps.

Definition 1. *(Robot Observation Data Frame) At each time step t, the robot data frame is represented as a tuple:*

$$\mathcal{D}_t = \left(\mathcal{I}_t^{RGB}, \mathcal{I}_t^{Stereo}, \mathbf{p}_t^{EE}, s_t^{Gripper}, \mathbf{f}_t\right)$$

where $\mathcal{I}_t^{RGB} = \{I_t^{RGB} \in \mathbb{R}^{H \times W \times 3}\}$ is a set of multi-view RGB images captured from monocular cameras, $\mathcal{I}_t^{Stereo} = \{I_t^{Stereo} \in \mathbb{R}^{2 \times H \times W \times 3}\}$ denotes a set of multi-view stereo image pair (left and right) if available, $\mathbf{p}_t^{EE} \in SE(3)$ is the 6-DoF end-effector pose and $s_t^{Gripper} \in \mathbb{R}$ denotes the scalar gripper state *such as gripper aperture, $\mathbf{f}_t \in \mathbb{R}^6$ is the force-torque vector from the end-effector sensor.*

The camera images are referred as exteroceptive sensing and the end-effector-related states belong to proprioceptive sensing.

Definition 2. *(Robot Trajectory) A trajectory $\mathcal{T}$ is defined as the temporally ordered sequence of observations $\mathcal{D}_{1:T}$ with a trajectory task language description l:*

$$\mathcal{T} = \{\mathcal{D}_{1:T}, l\}$$

Given a robot trajectory, Robo2VLM begin with *scene-interaction understanding*, applying semantic segmentation and manipulation phase classification to identify key segments (e.g., pre-grasp/approaching, contact, grasp, release). From these, we extract *keyframes* based on phase transitions, scene coverage, and visibility of objects or the robot across multiple camera views. We use manipulation domain knowledge to design *question prototype* to target core manipulation skills such as spatial relationship, goal conditions, and interaction understanding. Robo2VLM instantiates these prototypes on selected keyframes and transforms them into natural language multiple-choice questions via a *visual-language grounding* module that performs question conversion and spatial query projection.

A. Scene-Interaction Understanding

a) Embodied Scene Understanding: Given a task description in nature language and all images from different camera views, we first parse the language instruction using an off-the-shelf LLM such as Qwen 2.5 [2] to obtain {`target object`}, scene, task, and skill description. For the spatial understanding in manipulation, we need to know the relative direction and displacement between target object and gripper. From the proprioceptive data, we obtain the target object interaction point ground-truth from the robot trajectory data frames.

b) Manipulation Phase Segmentation: To segment robotic manipulation trajectories into semantically meaningful phases, we define a temporal phase classification function based on the sequence of end-effector poses, gripper aperture signals, and force-torque measurements: $\mathbf{p}^{\text{EE}}_{1:T}, s^{\text{Gripper}}_{1:T}, \mathbf{f}_{1:T}$. To align different types of gripper aperture, s^{Gripper}_t is normalized to $[0, 1]$, where 0 indicates fully open and 1 indicates fully closed. Let $s_t \in [0, 1]$ denote the normalized aperture at time t, and $\Delta s_t = s_t - s_{t-1}$ its temporal derivative. $\Delta s_t \approx 0$ denotes a small change within a tolerance margin ϵ, typically set to filter out noise. Let $\|\mathbf{f}_t\|$ be the force magnitude (if available). We introduce three threshold parameters: τ_g (grasp threshold), τ_c (closure threshold), and τ_f (force threshold for contact detection). Manipulation processes can be represented as a sequence of discrete phases, including approaching, stabilizing, contacting, releasing, and resetting or transitioning to subsequent actions. We denote the phase varible as $\Phi = \{\Phi_{\text{app}}, \Phi_{\text{stab}}, \Phi_{\text{cont}}, \Phi_{\text{rel}}, \Phi_{\text{reset}}, \Phi_{\text{trans}}\}$. Each timestep t is assigned a label $\phi_t \in \Phi$ according to the following temporal logic rules:

$$\phi_t = \begin{cases} \Phi_{\text{app}} & \text{if } s_t < \tau_g \wedge \Delta s_t < -\epsilon \\ \Phi_{\text{stab}} & \text{if } \phi_{t-1} = \Phi_{\text{app}} \wedge s_t < \tau_g \wedge |\Delta s_t| \leq \epsilon \\ \Phi_{\text{cont}} & \text{if } \phi_{t-1} = \Phi_{\text{stab}} \wedge s_t \geq \tau_c \wedge |\Delta s_t| \leq \epsilon \\ \Phi_{\text{rel}} & \text{if } \phi_{t-1} = \Phi_{\text{cont}} \wedge s_t \geq \tau_c \wedge \Delta s_t > \epsilon \\ \Phi_{\text{reset}} & \text{if } \phi_{t-1} = \Phi_{\text{rel}} \wedge s_t < \tau_g \wedge \Delta s_t > \epsilon \\ \Phi_{\text{trans}} & \text{otherwise} \end{cases}$$

The inclusion of force magnitude ensures that passive closure without external contact is not misclassified as active interaction. This multimodal phase labeling strategy captures both kinematic intent and physical contact, enabling robust segmentation of diverse manipulation behaviors.

To enforce a temporally coherent yet flexible phase progression, we define a partial order over the manipulation phases:

$$\Phi_{\text{app}} \prec \Phi_{\text{stab}} \prec \Phi_{\text{cont}} \prec \Phi_{\text{rel}} \prec \Phi_{\text{reset}} \rightarrow \Phi_{\text{app}}$$

This structure enforces unidirectional transitions along the phase chain, while allowing both phase skipping (e.g., directly from Φ_{app} to Φ_{cont}) and looping from the terminal phase Φ_{reset} back to the initial phase Φ_{app}, which is common in sequential manipulation routines. At each time step t, the phase label must satisfy $\phi_t \succeq \phi_{t-1}$, or $\phi_t = \Phi_{\text{app}}$ if $\phi_{t-1} = \Phi_{\text{reset}}$, ensuring temporal monotonicity or task repetition without reversal. The auxiliary state Φ_{trans} is used for ambiguous, missing, or conflicting observations where no confident assignment is possible. This symbolic, temporally-constrained model supports robust segmentation of complex manipulation behaviors under noisy or partially missing sensory input.

B. Visual Question Prototype

We design a set of *visual question prototypes*, each of which aligns with specific manipulation task completion required robot capabilities and anchors to distinct manipulation phases as illustrated in Table I. These prototypes are organized into three reasoning categories.

Spatial Reasoning focuses on the robot's understanding of object geometry, reachability, and spatial layout across viewpoints. Questions such as "Is the object reachable?" or "What's the relative direction between the gripper and the object?" are grounded in the early approach 🟥 and stabilize 🟨 stages. These rely on RGB, depth, stereo, and 3D gripper pose data, which together enable accurate localization and spatial inference across frames or views.

Goal-conditioned Reasoning probes the agent's high-level understanding of tasks, including goal inference, future action prediction, and overall task success. Questions such as "Is the task failed?", "What will the robot do next?", and "What is the robot's current action phase?" span multiple manipulation phases from approach 🟥 through reset 🟪. These require temporal context, pose estimation, and sometimes motion history, leveraging the multi-step evolution of the scene.

Interaction Reasoning focuses on physical interaction dynamics, such as grasp stability or the robot's current actuator state. These occur during stabilize 🟨, contact 🟩, and release 🟦

TABLE I: Categorization of visual reasoning questions for robotic manipulation, with manipulation phase (color-coded) and data modality context. 🟥 Approach, 🟨 Stabilize, 🟩 Contact, 🟦 Release, 🟪 Rest.

Capabilities	Question Prototype	Manip. Phase	Sensor Modality
	Spatial Reasoning		
Object State	Is the {`target object`} reachable by the robot?	🟥	I_t^{RGB}, D_t
Spatial Relationship	What's the relative direction in 3-D between end effector and {`target object`}?	🟥 🟨	I_t^{RGB}, $\mathbf{p}_t^{\mathrm{EE}}$
Scene Understanding	Which point is closer to the camera viewing the scene?	🟥 🟨	I_t^{RGB}, I_t^{Stereo}
Multiple View	Which point in the right-side image corresponds to the point in the left-side image?	🟥 🟨 🟦 🟪	I_t^{Stereo}
	Goal-conditioned Reasoning		
Task State-success	Has the robot successfully completed the task?	🟪	I_t^{RGB}
Task State-Goal	What is the goal configuration for {`interaction`}?	🟥 🟨 🟩 🟦	I_t^{RGB}, $\mathbf{p}_t^{\mathrm{EE}}$
Action Understanding	The robot is {`interaction`}. What is the robot's current action phase?	🟥 🟨 🟩 🟦 🟪	I_t^{RGB}, $\mathcal{T}_{1:t}$
Interaction Phase	What will the robot do next?	🟥 🟨 🟩 🟦	I_t^{RGB}, $\dot{\mathbf{p}}_t^{\mathrm{EE}}$
Trajectory Understanding	What task does this trajectory likely accomplish?	🟥	I_t^{RGB}, $\mathbf{p}_t^{\mathrm{EE}}$
	Interaction Reasoning		
Task State-grasp	Is this a stable grasp?	🟨 🟩 🟦	I_t^{RGB}, $\mathbf{f}_t$
Robot State	Is the robot gripper currently open?	🟨 🟩 🟦	I_t^{RGB}, s_t^{Gripper}

phases, and depend on RGB, tactile, or gripper aperture signals. For instance, the question "Is this a stable grasp?" may depend on contact force readings or inferred object displacement.

The ground truth of the questions are grounded by multiple sensor modality observations. We design the incorrect answers as part of the visual question prototypes. For example, in the scene understanding, we require the sampled points to be significantly different in depth from other points and from the depth sensor to account for sensor inaccuracy. In action understanding, the correct action arrow differs significantly from the distractor arrows by having a large angular separation in the projected 2D image. To detect guessing by hallucination, we randomly replace some correct answers with "None of Above" option.

C. *Keyframe Selection*

Given that raw robotic trajectories often contain hundreds of frames sampled at high frequency, using all frames is computationally expensive and can introduce redundancy due to minimal temporal variation. Moreover, many intermediate frames are visually or semantically uninformative for downstream reasoning tasks. To address this, we select a compact set of keyframes that retain essential semantic and visual cues while reducing redundancy and data volume. These keyframes are extracted from the multi-modal robot trajectory $\mathcal{T} = \{\mathcal{O}_t\}_{t=1}^{T}$ based on manipulation phase transition, scene coverage diversity and context visibility.

IV. Robo2VLM-1 Dataset

Open X-Embodiment and its datasets Open X-Embodiment [23] is major collaborative research initiative that aggregates robotic demonstration data collected from 22 different robot embodiments across 35 research labs worldwide, encompassing over 1 million trajectories covering more than 500 skills. Applying domain reweighting [24], we select a subset focusing on manipulation with real robot embodiments. In total, we use 13 datasets [28, 30–43] with a total of 176,139 trajectories. While most modalities are included in Open X-Embodiments release, we manually include modalities introduced by the original paper. For example, DROID dataset [30] includes camera calibration information and stereo depth. The detailed modality inclusion can be found in Table. II.

TABLE II: Trajectories and sensing modalities across datasets with a total of 176k trajectories. **# Traj**: number of trajectories; **Prop**: joint-state proprioception; **Dpth**: depth images; **GripAp**: gripper-aperture signal; **# VQA**: number of questions. ✓ denotes modality is available, ✗ denotes absent.

Dataset	# Traj	Prop	Dpth	GripAp	# VQA
DROID [30]	92k	✓	✓	✓	299k
Fractal [27]	73k	✓	✗	✓	267k
Kuka MM [33]	3k	✓	✓	✓	25k
Autolab [34]	896	✓	✓	✓	22k
Sirius [35]	600	✓	✗	✓	21k
MVP [36]	480	✓	✗	✓	8k
VINN [37]	435	✗	✗	✗	34
Fanuc [38]	415	✓	✗	✓	11k
TableTop [40]	110	✓	✗	✓	5k
VIOLA [41]	135	✓	✗	✓	8k
BUDS [42]	50	✓	✗	✓	6k
ROT [43]	14	✓	✗	✓	245

Robo2VLM for Open X-Embodiment We use Robo2VLM to process each robot trajectory from the Open X-Embodiment dataset by selecting and interpreting the scenes. The entire process takes 2935.7 GPU hours on Nvidia A100 GPUs. For each selected keyframe, Robo2VLM instantiates questions from embodied question templates resulting in the generation of a pool of over 3 million VQA items.

a) Robo2VLM-1 Curation: Inspired by data optimization paradigms such as domain reweighting in natural language processing [24] and robot policy learning [25], our curation process aims to balance the distribution of questions across diverse scene and task types. It selects a representative and high-quality subset of questions that effectively balances diversity across scenes, tasks, skills, and reasoning types, while ensuring clarity and unambiguous ground truth. In total, Robo2VLM-1 contains 684,710 questions, spanning 463 distinct real-world scenes, 3,396 unique robotic manipulation tasks, and 149 different manipulation skills.

V. Experiment

In this section, we sample 60k VQA from Robo2VLM-1 with a 50k training set (Robo2VLM-1-Train) and a 10k testing set (Robo2VLM-1-Test). We mainly study two research questions: (1) How does Robo2VLM-1-Train dataset improve the spatial and interaction reasoning capabilities of VLMs? and (2) How effectively does Robo2VLM-1-Test evaluate VLMs in these reasoning tasks?

Evaluation Setup We benchmark state-of-the-art open-source models in different configurations, including LLaVA, Llama 3.2 Vision, and Qwen2-VL/Qwen2.5-VL. Each model is evaluated under both zero-shot and Chain-of-Thought (CoT) prompting settings. For CoT, we follow the prompting strategy from [10] by appending the following instruction to the end of each question: *"Reason step by step about the answer, and show your work, for each step. Only after that, proceed to the final answer."* We run a simultaneous Llama-3.2-3B-Instruct to extract model outputs for final letter answer. We focus fine-tuning on language layers (both attention and MLP modules) while keeping vision layers frozen. For each configuration, we use random 2000 questions from the testing set. For consistency, all models are evaluated with a temperature of 0.7, a maximum completion token length of 4096, and overall context length of 10240. All models use their vision or vision instruct version with float16 quantization. All models are evaluated with 8 Nvidia A100 GPUs with 80GB memory. We use LoRA to fine-tune LLaVA 1.6 with rank 128 and alpha 256.

A. Benchmark with Robo2VLM-1

Table III presents a detailed comparison of vision–language foundation models on the Robo2VLM-1 benchmark, evaluated under both zero-shot and Chain-of-Thought (CoT) prompting conditions. The results reveal nuanced interactions across model architecture, scale, and reasoning strategy.

Cross-Model Performance: Evaluation data on Robo2VLM-1-test suggests that Qwen models has higher overall accuracy compared to other VLMs of the same configuration, which align with the observation from other VQA benchmarks such as [50, 51]. Qwen 2.5 VL-72B achieves the highest zero-shot accuracy at 37.76%, while Qwen 2.5 VL-32B achieves 41.30% overall accuracy in the CoT setting. Qwen models particularly excel in object-centric categories such as Object State, where Qwen 2.5 VL-72B reaches 85.00% (zero-shot) and 92.37% (CoT), and Interaction Phase (IP) (71.09% zero-shot, 74.09% CoT for 72B).

Impact of Model Scale. Zero-shot accuracy generally improves with model size — rising from 30.63% (Qwen 7B) to 37.76% (Qwen 72B). However, this trend does not hold in the CoT setting, where the 32B model surpasses the 72B model (41.30% vs. 39.52%). The observation aligns the official technical report of Qwen2.5[2] that the mathematical and problem-solving capabilities of Qwen2.5-VL-32B are further enhanced through reinforcement learning. LLaMA models display a different trend — while the 11B model outperforms the 90B version in zero-shot setting, the larger model benefits more under CoT prompting, suggesting that scaling may unlock latent capabilities only when paired with explicit reasoning support.

Effectiveness of CoT Prompting: CoT prompting generally enhances performance for both Qwen and LLaMA models. For example, Qwen 2.5 VL-7B improves from 30.63% to 34.82%, and LLaMA 3.2-90B increases from 28.60% to 30.45%. The most substantial gains are observed in Qwen 2.5 VL-32B, which improves from 37.68% to 41.30%. Results suggest that CoT benefits Task State–Success(from 55.08% to 60.43%), and Interaction Phase (from 63.80% to 71.35%). However, in the Spatial Relationship category, for example, Qwen 32B's accuracy drops from 21.85% to 18.82%, indicating that verbose reasoning chains may introduce noise in tasks requiring precise spatial localization.

B. Finetuning with Robo2VLM-1

We perform model finetuning experiment using Robo2VLM-1-train and evaluate on Robo2VLM-1-test. We increase the training data samples from 10k to 50k in finetuning. As depicted in Figure 2, increasing the fine-tuning data generally leads to notable performance enhancements across most VQA categories. Significant gains are observed in 'Object State' understanding, where accuracy improved from 29.34% to 80.24%. "Task State-success" also sees a substantial rise from 47.65% to 68.03%. Other categories demonstrating clear positive trends with more data. However, in some categories such as Spatial Relationship and Task State–Goal, fine-tuning with limited data (e.g., 10k) underperforms the no-finetuning baseline. This may be because the model has not yet seen enough task-specific examples to begin generalizing, or because the question formats in Robo2VLM-1 differ from those seen during pretraining, requiring adaptation time. In some categories, finetuning with Robo2VLM-1 does not improve the performance due to the reasoning capability limitation of the base model. This is also reflected in the fact that LLaVA shows performance degradation in CoT prompting in Table III. The "interaction phase" question requires the model to predict the next frame, demanding complex reasoning and making it a particularly challenging problem. This suggests that for complex tasks, the base model language performance is important for further improvement with Robo2VLM-1.

C. Comparison with Human Performance

We conducted a human evaluation covering all 11 categories defined in Table III. For each category, a human evaluator was asked to randomly answer questions from Robo2VLM-1-test. We use the average success rate as a reference for comparison with three models—LLaVA 1.6-7B, LLaVA 1.6-7B-Finetuned, and Qwen 2.5 VL-32B—CoT on the same set of categories as shown in Figure 3. Qwen 2.5 VL-32B—CoT

TABLE III: Performance Comparison of Multimodal Foundation Models on OpenX-VQA Benchmark Categories (%). Upper part: zero-shot. Lower part: with CoT prompting.

		Spatial Reasoning					Goal Reasoning			Interaction Reasoning		
Model	**Overall**	**RS**	**OS**	**SR**	**SU**	**MV**	**TS-G**	**TS-S**	**TS-GL**	**AU**	**IP**	**TU**
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Zero-Shot												
LLaVA 1.5-7B	21.58	35.32	23.87	16.08	17.78	17.50	31.82	23.79	19.03	20.30	21.74	22.37
LLaVA 1.6 Mistral-7B	24.09	30.31	35.13	19.42	20.24	**19.29**	34.20	30.77	**19.52**	18.67	20.70	22.83
LLaVA 1.6-34B	24.94	26.66	29.75	21.47	23.18	17.86	29.19	29.40	17.90	19.49	36.98	30.59
Llama 3.2-90B	28.60	31.94	55.87	18.51	26.61	16.43	28.23	35.27	8.06	18.13	51.56	49.77
Qwen 2.5 VL-7B	30.63	41.68	55.63	21.55	24.38	17.32	33.01	42.57	7.82	25.71	46.61	39.73
Qwen 2.5 VL-32B	37.68	**49.39**	71.37	21.85	**28.53**	17.50	**34.21**	**55.08**	12.90	30.45	63.80	49.32
Qwen 2.5 VL-72B	**37.76**	38.84	**85.00**	**22.31**	28.23	15.71	28.47	51.89	10.08	**33.96**	**71.09**	**54.79**
CoT Reasoning												
LLaVA 1.5-7B	21.61	28.28	21.00	17.37	20.90	18.93	25.36	24.19	**21.53**	21.24	20.31	20.09
LLaVA 1.6 Mistral-7B	24.05	27.60	38.87	17.15	20.18	**22.32**	25.84	28.03	18.47	18.40	30.60	29.68
LLaVA 1.6-34B	23.49	20.43	31.00	21.24	22.88	20.36	18.18	26.14	16.77	21.79	35.16	26.94
Llama 3.2-90B	30.45	32.34	79.87	13.35	26.37	18.57	29.90	29.14	14.27	19.76	59.24	44.75
Qwen 2.5 VL-7B	34.82	38.02	90.00	**21.78**	23.30	16.79	**36.84**	46.48	18.39	28.15	42.71	36.99
Qwen 2.5 VL-32B	**41.30**	**48.85**	90.50	18.82	29.19	19.82	35.17	**60.43**	18.71	32.21	71.35	**49.32**
Qwen 2.5 VL-72B	39.52	44.79	**92.37**	18.36	**29.73**	13.39	29.19	55.28	13.15	**36.13**	**74.09**	46.12

Category Abbreviations: **Spatial Reasoning:** RS: Robot State, OS: Object State, SR: Spatial Relationship, SU: Scene Understanding, MV: Multiple View. **Goal Reasoning:** TS-G: Task State-grasp, TS-S: Task State-success, TS-GL: Task State-goal. **Interaction Reasoning:** AU: Action Understanding, IP: Interaction Phase, TU: Trajectory Understanding.

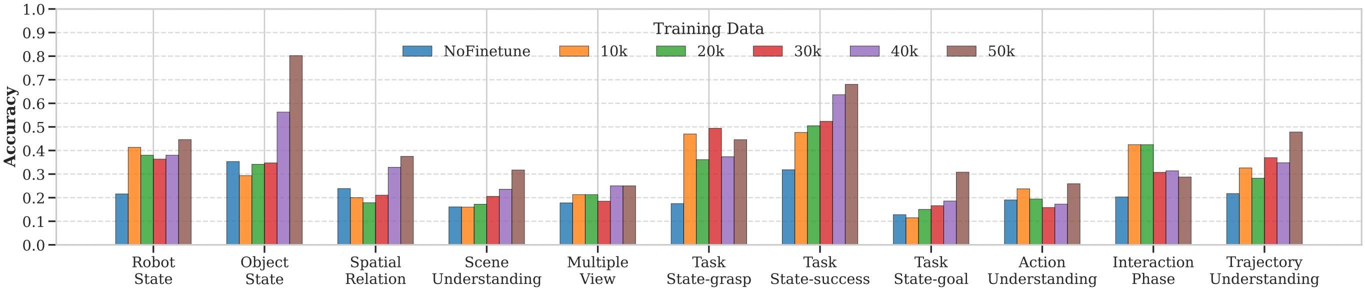


Fig. 2: **Fine-tuning LLaVA 1.6 with increasing training data of Robo2VLM-1** from 10k to 50k VQA items. Accuracy improvements almost all categories compared to no fine-tuning.

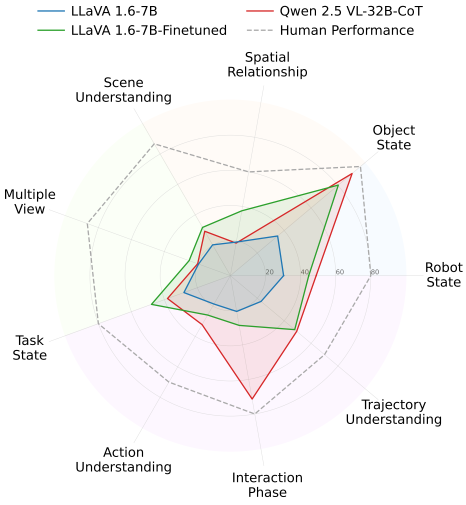


Fig. 3: Comparison of human performance to different multimodal foundation models.

achieves near human accuracy, with 90.5% in Object State compared to 96.7% for humans, and 71.35% in Interaction Phase versus the human score of 80.0%. In more complex spatial reasoning tasks such as Spatial Relationship, where human achieves 60.0% accuracy, the best model (LLaVa 1.6-7B, finetuned) reaches only 19.42%. This may suggest that even if observing from multiple views, a monocular image may lack the full depth information needed to accurately determine the spatial relationship. Furthermore, finetuning enhances model performance. LLaVA 1.6-7B finetuned on the Robo2VLM-1 training dataset shows consistent improvements across multiple categories, particularly in Task State, Object State, and Trajectory Understanding, compared to its non-finetuned LLaVA 1.6-7B. These findings demonstrate the potential Robo2VLM-1 in studying and narrowing the gap between model and human performance in spatial and task reasoning.

References

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 8748–8763.

[2] Q. Team, *Qwen2.5: A party of foundation models*, Sep. 2024.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[4] Anthropic, *Claude 3.5 Sonnet*, https://www.anthropic.com/news/claude-3-5-sonnet, Jun. 2024.

[5] OpenAI, *GPT-4o System Card*, https://openai.com/index/gpt-4o-system-card/, Aug. 2024.

[6] K. Kavukcuoglu, *Gemini 2.5: Our most intelligent AI model*, https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, Mar. 2025.

[7] B. Chen *et al.*, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 14 455–14 465.

[8] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, *Prismatic vlms: Investigating the design space of visually-conditioned language models*, 2024. arXiv: 2402.07865 [cs.CV].

[9] M. Kim *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[10] G. R. Team, S. Abeyruwan, *et al.*, *Gemini robotics: Bringing ai into the physical world*, 2025. arXiv: 2503.20020 [cs.RO].

[11] L. X. Shi *et al.*, *Hi robot: Open-ended instruction following with hierarchical vision-language-action models*, 2025. arXiv: 2502.19417 [cs.RO].

[12] M. M. Islam, A. Gladstone, R. Islam, and T. Iqbal, "EQA-MX: Embodied question answering using multimodal expression," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[13] R. Yang *et al.*, "EMBODIEDBENCH: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents," *arXiv preprint arXiv:2502.09560*, 2025.

[14] M. Li *et al.*, "Embodied agent interface: Benchmarking LLMs for embodied decision making," in *NeurIPS 2024 Track on Datasets and Benchmarks*, 2024.

[15] M. Shridhar *et al.*, "ALFRED: A benchmark for interpreting grounded instructions for household robots," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] A. Szot, E. Coumans, A. Collett, and et al., "Habitat 2.0: Training home assistants to rearrange their habitat," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[17] E. Kolve, R. Mottaghi, D. Gordon, and et al., "AI2-THOR: An interactive 3d environment for visual AI," *arXiv preprint arXiv:1712.05474*, 2017.

[18] P. Sermanet *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," *arXiv preprint arXiv:2311.00899*, 2023.

[19] Y. Ji *et al.*, "Robobrain: A unified brain model for robotic manipulation from abstract to concrete," *CVPR*, 2025.

[20] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[21] Octo Model Team *et al.*, *Octo: An open-source generalist robot policy*, https://octo-models.github.io, 2023.

[22] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 23.

[23] O. X.-E. Collaboration *et al.*, *Open X-Embodiment: Robotic learning datasets and RT-X models*, https://arxiv.org/abs/2310.08864, 2023.

[24] S. M. Xie *et al.*, "Doremi: Optimizing data mixtures speeds up language model pretraining," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69 798–69 818, 2023.

[25] J. Hejna, C. A. Bhateja, Y. Jiang, K. Pertsch, and D. Sadigh, "Remix: Optimizing data mixtures for large scale imitation learning," in *Proceedings of The 8th Conference on Robot Learning*, P. Agrawal, O. Kroemer, and W. Burgard, Eds., ser. Proceedings of Machine Learning Research, vol. 270, PMLR, Jun. 2025, pp. 145–164.

[26] A. Sharma *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2310.01894*, 2023.

[27] A. Brohan *et al.*, "Rt-1: Robotics transformer for real-world control at scale," 2023. arXiv: 2212.06817 [cs.RO].

[28] A. Brohan *et al.*, *Rt-2: Vision-language-action models transfer web knowledge to robotic control*, 2023. arXiv: 2307.15818 [cs.RO].

[29] K. Black *et al.*, *π_0: A vision-language-action flow model for general robot control*, https://physicalintelligence.company/blog/pi0, 2024.

[30] A. Khazatsky, K. Pertsch, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," 2024.

[31] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard, "Latent plans for task agnostic offline reinforcement learning," 2022.

[32] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[33] M. A. Lee *et al.*, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[34] L. Y. Chen, S. Adebola, and K. Goldberg, *Berkeley UR5 demonstration dataset*, https://sites.google.com/view/berkeley-ur5/home.

[35] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," in *Robotics: Science and Systems (RSS)*, 2023.

[36] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *CoRL*, 2022.

[37] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, *The surprising effectiveness of representation learning for visual imitation*, 2021. arXiv: 2112.01511 [cs.RO].

[38] X. Zhu, R. Tian, C. Xu, M. Ding, W. Zhan, and M. Tomizuka, "Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot," 2023.

[39] Y. Zhou, S. Sonawani, M. Phielipp, S. Stepputtis, and H. Amor, "Modularity through attention: Efficient training and transfer of language-conditioned policies for robot manipulation," in *Conference on Robot Learning*, PMLR, 2023, pp. 1684–1695.

[40] Y. Zhou, S. Sonawani, M. Phielipp, H. Ben Amor, and S. Stepputtis, "Learning modular language-conditioned robot policies through attention," *Autonomous Robots*, pp. 1–21, 2023.

[41] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," *6th Annual Conference on Robot Learning (CoRL)*, 2022.

[42] Y. Zhu, P. Stone, and Y. Zhu, "Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4126–4133, 2022.

[43] S. Haldar, V. Mathur, D. Yarats, and L. Pinto, “Watch and match: Supercharging imitation with regularized optimal transport,” in *Conference on Robot Learning*, PMLR, 2023, pp. 32–43.

[44] S. Antol *et al.*, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[45] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[46] J. H. Lee, M. Kerzel, K. Ahrens, C. Weber, and S. Wermter, “What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning,” *arXiv preprint arXiv:2205.02671*, 2022.

[47] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, “Embodied question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[48] P. Anderson *et al.*, “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[49] H. I. Christensen and G. D. Hager, “Sensing and estimation,” in *Springer Handbook of Robotics*, ser. Springer Handbooks, B. Siciliano and O. Khatib, Eds., Springer, 2016, pp. 91–112.

[50] P. Lu *et al.*, *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts*, 2024. arXiv: 2310.02255 [cs.CV].

[51] D. Hendrycks *et al.*, *Measuring massive multitask language understanding*, 2021. arXiv: 2009.03300 [cs.CY].