Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

RoboVista: Evaluating Vision Language Models for Diverse Robot Applications

Shuangyu Xie[1,*] Kaiyuan Chen[1,*] Ziyang Chen[1], Simeon Adebola[1], Yixuan Huang[2], Zehan Ma[1], Tianshuang Qiu[1], Wentao Yuan[3], Dhruv Shah[2,3], Pannag R Sanketi[3], Ken Goldberg[1]

[1] University of California, Berkeley [2] Princeton University [3] Google DeepMind * equal contribution

Fig. 1: **RoboVista Overview.** To support future robot applications, RoboVista presents fine-grained spatial understanding and embodied decision-making challenges for Vision–Language Models (VLMs). Grounded in 6 robot application domains and 39 diverse tasks, RoboVista is an expert-annotated Visual Question Answering (VQA) dataset emphasizing variable robot embodiments (left), interactions with deformable objects and complex and cluttered scenes (middle), and long-horizon contextual understanding (right).

***Abstract*—Diverse applications for robotics, such as industry and agriculture, require robots to operate across various embodiments, changing visual conditions, and complex planning. Vision–Language Models (VLMs) offer a promising foundation for general-purpose and interpretable robotic reasoning. Aligning VLMs with diverse robot applications requires a *modular* understanding of the individual decision components that underlie robotic behavior. Capturing such structure is challenging for conventional robot benchmarks that are primarily based on teleoperated, end-to-end datasets. We propose Robot Question Answering (RQA), a modular evaluation framework and RoboVista, a benchmark curated from real robotic systems, research papers, and expert annotations. RoboVista contains 474 Visual Question Answering (VQA) instances with human annotated reasoning and covers 39 unique task types in agricultural, industrial, domestic, surgical robotics, autonomous driving, and open robot datasets. Experiments on RoboVista show that state-of-the-art VLMs exhibit substantial gaps. Physical robot experiments suggest strong correlation between RoboVista performance and real-world task execution. https://berkeleyautomation.github.io/robovista**

I. Introduction

Deploying robots in real-world domains such as industrial automation [1, 2], agriculture [3, 4], and surgery [5, 6] requires a general-purpose interface that can reason over diverse decisions, embodiments, and situations. Vision-Language Models (VLMs) [7, 8] are a potential backbone for these systems, as their language-driven reasoning has shown strong capabilities in complex domains such as software engineering [9] and has shown early success in robot manipulation [10–12]. However, to support real robot tasks, VLMs must deliver consistent, spatially grounded decision-making: for example, on an industrial gasket assembly line, a VLM must accurately judge whether a gasket is properly seated and choose the correct next action (re-seat or fasten). Toward this goal, we present RoboVista, a high-quality benchmark that covers diverse robots application scenarios (as shown in Fig. 1).

A primary way to evaluate this capability is Visual Question Answering (VQA) [13–18], where a model is given images or videos and must answer task-relevant questions that probe scene understanding and action selection. Existing efforts, such as Robo2VLM [19] and RoboBrain [20], largely draw from imitation learning datasets such as Open X-Embodiment [21], and have shown success in improving spatial capabilities. However, many existing real-world robotic applications are fundamentally *modular*: complex behaviors are decomposed into task-level decisions (e.g., task sequencing, action planning, and recovery) that are hard to capture by end-to-end robot trajectories. This motivates extending evaluation to these modular decisions with broader application domains, especially where public data is scarce, and many critical decisions are handled by analytical pipelines.

We propose RQA, a modular evaluation framework for Vision–Language Models (VLMs) that systematically constructs high-quality, robot-centric VQAs. RQA decomposes diverse robot applications using standard robotic abstractions, and unifies human expert annotation, algorithmic task execution, and automated question construction within a shared Robot-VQA interface [22]. RQA provides a structured interface for evaluating VLM capabilities by formalizing both a module-level problem abstraction and a principled VQA construction process.

We introduce RoboVista, an expert-annotated robot-centric VQA benchmark. It contains 474 multiple-choice questions covering 39 distinct robot task types spanning surgical, agricultural, industrial, domestic, autonomous driving, and open robot datasets. Each question is grounded in robot-visible or onboard

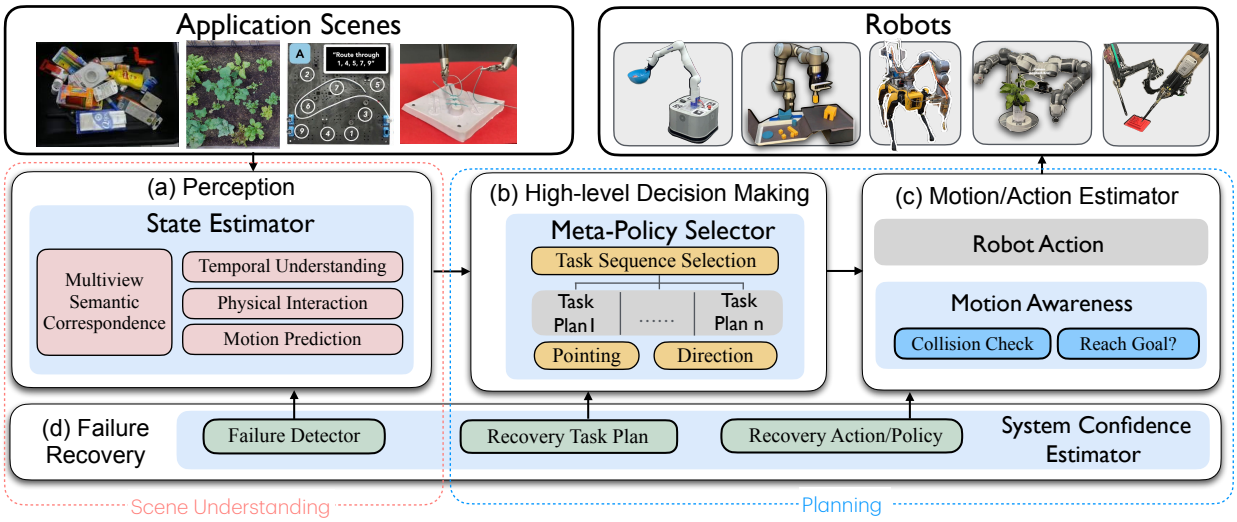


Fig. 2: **Module abstraction of diverse robot applications pipelines**. The figure illustrates a modular view of real-world robot systems, spanning diverse application scenes and robot embodiments. Robot operation can be decomposed into four functional layers in perception, high-level decision making, motion / action estimation and failure recovery.

visual observations and paired with detailed human reasoning explanations. To ensure realism and consistency across domains, all questions are constructed and verified by human domain experts familiar with the underlying robotic algorithms and task constraints.

We perform comprehensive evaluation of state-of-the-art VLMs on RoboVista with varying techniques, such as Chain-of-Thought (CoT) [23] and In-Context Learning (ICL) [24, 25]. Evaluation data suggests there is a substantial gap for VLM in many robot application domains. We evaluate using a physical bi-manual spatial estimation task that uses domain-specific object geometry as priors. We further evaluate closed-loop VLM-guided surgical knot tying. Evaluation data suggests that RoboVista has strong correlation to task performance and the degradations in RoboVista performance significantly reduces long-horizon task success.

In this paper, we make the following contributions: (1) RQA: a modularized framework for constructing Visual Question Answering for VLMs from expert annotation, robot demonstration and execution; (2) RoboVista, a benchmark for VLMs spanning diverse robot applications, scenes, and embodiments; (3) a comprehensive evaluation of state-of-the-art VLMs on RoboVista, including cross-domain analysis, training and complementary physical-robot validation.

II. Related Work

Modular Robot System Design. Modular robot system design [22] has been the dominant paradigm for field and service robotics. In this paradigm, complex robot behaviors are decomposed into well-defined functional tasks, such as perception, state estimation, motion reasoning, decision making, and control, that can be implemented and analyzed independently . This paradigm has enabled successful systems across a wide range of applications, including surgical robotics [5, 26], agricultural systems [27–29], industrial automation [2, 30], and home robotics [31]. In many real-world domains, fully end-to-end data-driven approaches remain difficult to deploy due to the high cost and limited availability of training data. For example, surgical data must be obtained from real clinical procedures under strict safety, privacy, and regulatory requirements [5]. As a result, modularized pipelines continue to provide a practical and robust solution when data scaling itself is a hard problem, leveraging structured task decomposition and domain expertise.

VLMs in Robotic Systems. Recent advances in Large Language Models (LLMs) and VLMs have demonstrated strong capabilities in task reasoning, spatial understanding, and grounded decision-making for robotic systems [32, 33]. Increasingly, VLMs serve as the representational backbone of Vision–Language–Action (VLA) policies. Systems such as OpenVLA [34] and π_0[35] leverage pre-trained VLM representations to directly map visual observations and language goals to continuous robot actions. Beyond end-to-end control, VLMs are widely used for high-level task planning. Works such as Text2Motion[36] and PaLM-E [37] integrate language or multimodal models into task-and-motion planning frameworks, using them as semantic planners that translate language instructions and multimodal observations into executable, long-horizon plans. Related approaches generate code or programmatic representations as intermediate planning interfaces [11, 38–40], enabling compositional reasoning over perception, control, and symbolic constraints. More recent systems adopt modularized architectures in which VLMs orchestrate or invoke specialized perception, planning, and control components [41], combining the flexibility of foundation models with the structure of classical robotics pipelines. Despite this progress, it remains unclear whether the capabilities demonstrated by current VLMs are sufficient to support the full spectrum of robotic decision-making in real-world settings. Systematic evaluation is therefore required to reveal these gaps.

VQA as an Evaluation Paradigm. VQA has emerged as a primary interface for evaluating and benchmarking the capabilities of VLMs [13, 15, 16]. Early VQA benchmarks primarily assess general visual and linguistic abilities, such as commonsense knowledge (MMMU, MMMU-Pro [17, 18]), chart and diagram

understanding (ChartQA [42]), and mathematical reasoning (MathVista [43]).

Recent embodied VQA benchmarks evaluate VLMs on tasks like visual navigation and long-horizon planning [44–46], with newer works like ERQA [10], Robo2VLM [19], and RoboMind [47] incorporating richer embodiment and stronger task grounding. While frameworks such as Robo2VLM [19] and RoboBrain [20] leverage large-scale robot demonstrations to generate massive question-answer datasets for model post-training , RoboVista serves a complementary purpose. By designing VQA to capture perception-planning-control procedural methods across diverse applications , we enable evaluation in underrepresented domains—such as agriculture, surgery, and industrial robotics—that natively lack the massive trajectory corpora required by prior efforts.

Meanwhile, the focus of RoboVista on fine-grained diagnostic quality aligns with a broader shift in the language modeling community toward small but challenging benchmarks, such as TurnaboutLLM [48], MuSR [49], and FrontierMath [50], that prioritize reasoning depth over raw scale. We consider our scale a deliberate design choice rather than a limitation by ensuring equal weight during evaluation and avoid over-representing data-rich tabletop scenarios. Rather than removing domain expertise, RoboVista provides a structured framework to integrate critical, modular decision points into robotic VQA.

III. RQA Framework

We propose Robot Question Answering (RQA), a modular and unified framework that converts decision points from robot systems into a robot-centric VQA formulation. We first introduce a module-level abstraction of robot systems, then define the Robot-VQA data structure, describe the RQA construction process, and finally illustrate the framework with concrete case studies.

A. Module Abstraction

As illustrated in Fig. 2, RQA mirrors modular robot system pipelines [22], decomposing robot behavior into perception, decision making, motion execution, and recovery. Specifically, perception modules estimate task-relevant state from robot-centric visual observations, including object geometry, spatial relationships, and physical interactions. High-level decision making focuses on grounded task reasoning, requiring models to select goals, sequence actions, and determine appropriate next steps under task constraints. Motion and action awareness evaluates whether candidate actions are physically feasible, accounting for reachability, collision avoidance, and closed-loop visual feedback during execution. To handle failures that arise across different modules, recovery and robustness reasoning assess the ability to detect execution failures and infer corrective actions based on observed outcomes. Together, these stages provide a structured yet flexible abstraction for expressing diverse robot reasoning problems within a unified Robot-VQA formulation.

To facilitate the construction of Robot-VQA, we build upon established formalisms in perception [51], planning [52, 53], and operational control [54]. While individual abstractions exist for specific sub-problems, the general nature of Robot-VQA requires a unified framework. Inspired by the compositional logic of Neural Module Networks [55], we design a module-level problem abstraction where any functional block in the pipeline (Fig. 2) can be described by a tuple $\mathcal{M}$ with detailed explanations

$$\mathcal{M} = (\mathcal{E}, \mathcal{X}, \mathcal{U}, \mathcal{C}), \tag{1}$$

- $\mathcal{E}$ **Embodiment Setup** defines the robot's physical configuration and sensing capabilities. Following the configuration space approach [56], this includes kinematic structures, actuation limits, and sensor modalities that dictate the robot's interaction with the environment.
- $\mathcal{X}$ **State Space** represents the latent and observed variables of the robot and environment. Based on probabilistic robotics frameworks [57], this includes joint configurations, object poses, and task-relevant semantic attributes necessary for state estimation.
- $\mathcal{U}$ **Task Output Space** defines the decision or control manifold, such as discrete actions or motion parameters. The module implements produce high-level task decisions or operational space commands [54].
- $\mathcal{C}$ **Constraints** encode feasibility conditions (e.g., collision avoidance, joint limits) that restrict the valid output space and enforce geometric and kinematic safety requirements [58].

B. Robot-VQA Data Structure

Robot-VQA inherits from standard VQA while imposing robot-centric constraints to ensure embodied grounding.

Definition 1. *(VQA Instance with Reasoning) Let $\mathbb{V}$, $\mathbb{Q}$, $\mathbb{A}$, and $\mathbb{R}$ denote the spaces of visual inputs, questions, answers, and textual rationales, respectively. A VQA instance with reasoning is a 5-tuple $\mathcal{Q} = (\mathcal{V}, q, a^\star, \mathcal{A}, r) \in \mathbb{V} \times \mathbb{Q} \times \mathbb{A} \times 2^{\mathbb{A}} \times \mathbb{R}$, where $\mathcal{V}$ is the visual input, q the question, $a^\star$ the ground-truth answer, $\mathcal{A}$ the candidate answer set containing $a^\star$, and r a textual rationale justifying $a^\star$. We denote the set of all such instances by $\mathbb{S}_{VQA}$.*

Definition 2. *(Robot-Centric Visual Input) We define two robot-centric subsets of $\mathbb{V}$ such that robot is visible or image from onboard camera.*

$$\mathbb{V}_{rv} := \{\mathcal{V} \in \mathbb{V} : \mathrm{RobVis}(\mathcal{V})\}, \mathbb{V}_{ob} := \{\mathcal{V} \in \mathbb{V} : \mathrm{Onboard}(\mathcal{V})\}.$$

Definition 3. *(Robot-VQA Instance) A Robot-VQA instance is a subset of $\mathbb{S}_{VQA}$ whose visual input is grounded in the robot's embodied experience:*

$$\mathbb{S}_{\textit{Robot-VQA}} := \{\mathcal{Q} = (\mathcal{V}, q, a^\star, \mathcal{A}, r) \in \mathbb{S}_{VQA} : \mathcal{V} \in \mathbb{V}_{rv} \cup \mathbb{V}_{ob}\}.$$

C. RQA Design

With the formal definition of module abstraction $\mathcal{M}$ and Robot-VQA instance $\mathbb{S}_{\text{Robot-VQA}}$, the key question of RQA framework is how to achieve the conversion from $\mathcal{M}$ to $\mathbb{S}_{\text{Robot-VQA}}$. Concretely, RQA defines a structured mapping from a module specification to a Robot-VQA representation:

$$(\mathcal{E}, \mathcal{X}, \mathcal{U}, \mathcal{C}) \longrightarrow (\mathcal{V}, q, a^\star, \mathcal{A}, r), \tag{2}$$

where the module's latent state and embodiment are reflected through embodied visual observations, and its decision semantics are expressed through structured question–answering.

However, achieving robot-centric, modular VQA evaluation presents the following challenges for design considerations: (1) diverse robot applications rely on domain-specific modular pipelines, making it infeasible to provide a generic or fully automated construction procedure; (2) many decisions further depend on embodiment-specific or non-visual state (e.g., kinematics, contact, reachability) and sometimes visual data is subject to occlusions, limited viewpoints, and clutter; (3) designing plausible but incorrect answer choices is challenging, as each option must be algorithmically grounded and reflect a feasible outcome of the underlying module; (4) visual data and questions in certain domains and tasks are substantially harder to curate than in others.

To address these challenges, RQA adheres to the following design principles, each directly aligned with the challenges above: (1) robot VQAs are decomposed and structured by domain experts familiar with the robot application system; (2) each question is designed to be answerable with provided robot-centric visual and language context alone by the domain expert; (3) all questions and answer choices admit a single, well-defined correct answer that can be verified against the original module outcome; (4) the question curation procedure intentionally limits redundancy within each question and task type and prioritizes coverage across diverse robot tasks and questions.

Under this formulation, RQA convert modules to Robot-VQA as following. The visual input $\mathcal{V}$ consists of robot-centric images or image sequences that capture the task-relevant scene and robot configuration. These observations implicitly encode the underlying system state $\mathcal{X}$ used by the module, without explicitly exposing structured state variables in the VQA formulation. The question q describes the task-level decision associated with the module and specifies the objective to be reasoned about. When necessary, it also provides additional context about the embodiment setup $\mathcal{E}$ or task conditions that may not be visually explicit. The ground-truth answer $a^\star$ corresponds to a valid module output, i.e., an element of the task output space $\mathcal{U}$. For evaluation, the output space is discretized into a finite candidate set $\mathcal{A} \subseteq \mathbb{A}$ that includes $a^\star$ as well as plausible alternative outputs. Finally, embodiment, geometric, and kinematic constraints $\mathcal{C}$ governing the feasibility of module outputs are either observable from the visual input $\mathcal{V}$ or explicitly stated in the question q. The reasoning r provides a textual justification explaining why $a^\star$ satisfies these constraints while alternative candidates do not.

D. Case Studies: Instantiating RQA Across Domains

We present two representative case studies for applying RQA designs: (1) ambidextrous bin picking and (2) surgical knot tying, to illustrate how Robot Question Answering (RQA) systematically exposes decision points from diverse robotic pipelines as Robot-VQA instances. Despite differences in embodiment, contact dynamics, and task horizons, both systems can be decomposed into a shared set of functional modules aligned with Fig. 2.

Robot Grasping with Ambidextrous Robot Dex-Net 4.0 [59] is a representative work for addressing the ambidextrous bin-picking problem, where a bimanual robot clears a cluttered bin of heterogeneous objects using two complementary end-effectors: a suction gripper and a parallel-jaw gripper. The system samples a large set of candidate grasps over the visible object surfaces and evaluates each candidate using a Grasp Quality Convolutional Neural Network (GQ-CNN), which predicts a grasp robustness score under uncertainty. The robot then selects both the gripper type and grasp configuration that jointly maximize expected grasp success. A set of example Robot-VQA questions are shown in Fig. 3 (top).

Knot Tying with Surgery Robots Surgical knot tying is a long-horizon, contact-rich manipulation task that requires precise spatial reasoning, sequential decision making, fine-grained motion awareness, and continuous monitoring of execution outcomes. We frame knot tying on the da Vinci Research Kit (dVRK) [74] as a sequence of perception-grounded decision points. A set of example Robot-VQA questions are shown in Fig. 3 (bottom).

IV. The RoboVista Benchmark

We introduce *RoboVista*, a high quality robot-centric benchmark for VLMs constructed by RQA framework. RoboVista consists of 474 expert-annotated, multiple-choice Robot-VQA questions, each paired with 5 candidate answers and detailed reasoning explanations annotated by human expert. All questions are grounded in robot-centric visual observations and are curated to reflect decision points from real robotic applications. The visual data of RoboVista is curated from six application domains, industrial manufacturing, agriculture, domestic robotics, surgical robotics, autonomous driving and open robot datasets. It covers 39 distinct robotics task types, including bin picking, surgical knot tying, weed removal, cable routing, assembly, and navigation from 18 peer-reviewed conference and journal publications. This also includes questions from open robot datasets, DROID [73], Open X-Embodiment [21] and Agibot [72] with questions based on Robo2VLM framework [19]. To ensure annotation quality and task realism, all questions are annotated and verified by domain experts who are familiar with the underlying robotic algorithms and executions. All annotators are graduate-level and above, with more than half of the annotators holding Ph.D. degrees in Robotics.

The construction of RoboVista follows a three-stage data curation pipeline designed to ensure task realism, visual grounding, and cross-domain consistency.

Data Collection. First, we survey a broad range of robotics datasets and peer-reviewed publications across industrial, agricultural, domestic, surgical, and driving domains. From these sources, we identify representative robot application scenarios that expose meaningful perception, decision-making, motion, and recovery challenges. We then extract robot-centric visual data, including single images and short image sequences captured from onboard or robot-visible viewpoints, and group them into candidate task episodes.

Question Construction. In the second stage, we invite human domain experts (four postdoctoral researchers and three students

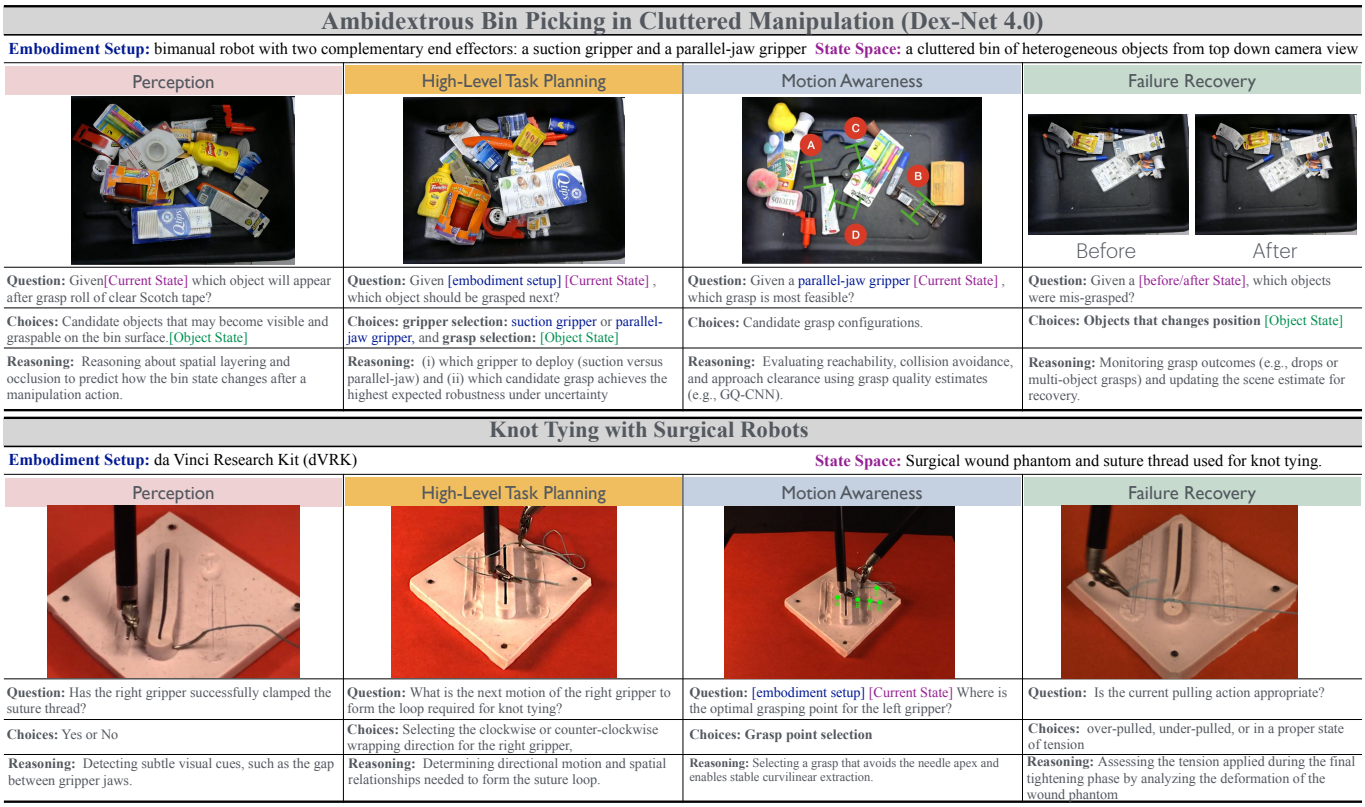


Fig. 3: **RQA case studies across domains.** This figure provide two example on how the RQA framework maps continuous physical robot states into modular, discrete VQA pairs that correspond to critical real-world decision points, such as tool alignment and tension management. Top: Ambidextrous bin picking with Dex-Net [59] Bottom: Surgical knot tying [5]. In both cases, robotic decision points are decomposed into perception, high-level planning, motion awareness, and failure recovery with RQA.

with master degree) to manually construct VQA questions from the selected episodes. Each question is designed to correspond to a concrete decision point within a modular robotics pipeline, such as grasp selection, next-action prediction, motion feasibility checking, or failure diagnosis. All questions are formulated as multiple-choice with five options, ensuring that each correct answer is visually grounded and operationally meaningful for robot execution. Annotators also provide a natural language explanation to justify the correct choice and clarify task intent.

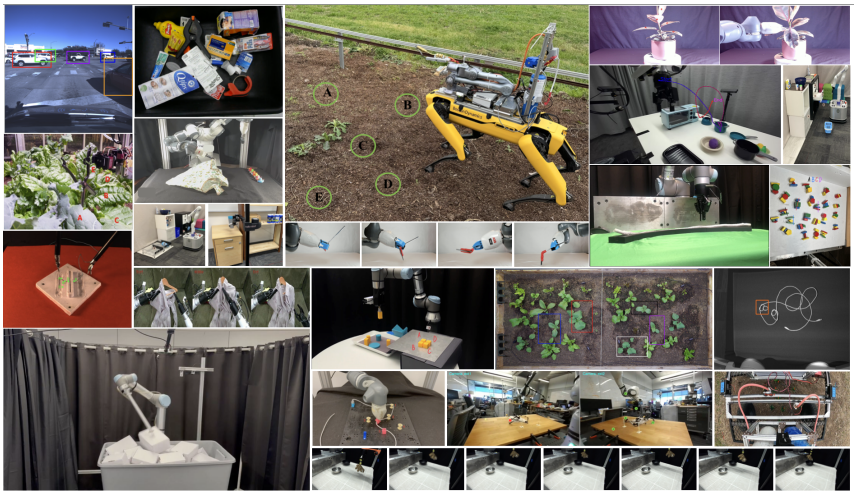
Quality Control. To ensure high annotation quality and consistency, all questions undergo a multi-pass review process. Each VQA instance is independently verified by at least one additional annotator with robotics expertise, who checks visual grounding, answer correctness, and linguistic ambiguity. Questions that admit multiple plausible answers, rely on non-visual cues, or lack clear task relevance are revised or discarded. We further enforce consistency across domains by aligning question types and difficulties. This process results in a curated benchmark that balances coverage, difficulty, and realism while minimizing annotation noise.

V. Experiments

In this section, we benchmark state-of-the-art open-source models in different configurations, including different sizes of Qwen2.5-VL [75] and Qwen3-VL [76]. For VLMs that are trained for robotics data, we benchmark Robo2VLM-ER [19], RoboBrain 2.5 [77]. For RoboBrain 2.5, we include a generic model for VQA and a model for spatial understanding (NV). Robo2VLM-ER uses Qwen2.5-VL-7B as base model, and RoboBrain 2.5 uses Qwen3-VL-8B as base model. For closed-source VLMs, we benchmark GPT-4o, GPT-5 (gpt-5-2025-08-07) and Gemini 2.5 Pro [78]. GPT-5 and Gemini 2.5 Pro are released at similar time frame. We also include results from Qwen-3-8B without any visual input to show how much linguistic hints provided in choice designs. All models are evaluated with a temperature of 0.7, a maximum completion token length of 4096, and overall context length of 10240.

A. Results on Zero-Shot Performance

Differences in VLM Models Table I summarizes zero-shot accuracy across all questions and across six domains. RoboVista overall poses non-trivial challenges. Even the best-performing models fall substantially short of perfect performance. Closed-source models achieve the strongest aggregate performance, with Gemini 2.5 Pro obtaining the highest overall accuracy (56.5%). Among open-source models, performance improves consistently with model scale, with Qwen 3-235B-A22B reaching 51.3% overall accuracy. Notably, robotics-specialized models such as RoboBrain 2.5-8B outperform similarly sized general-purpose VLMs on several domains, such as Driving and Surgery. However, these gains are not uniform across domains: for example, RoboBrain 2.5-8B-NV improves performance in domestic and surgical robotics tasks but underperforms in agricultural scenes.



Domain	Task Description	Perception	Planning	Total
Agriculture	Robot Gardening [60]	18	0	18
	Plant Inspection [29]	12	4	16
	Weed removal [28, 61, 62]	19	9	28
Driving	Self-Driving in Mcity [63]	9	11	20
Domestic	Garment manipulation [31]	1	2	3
	Domestic Tidying [64, 65]	30	19	49
Industrial	1D deformable [2, 66]	39	18	57
	3D deformable [67, 68]	7	8	15
	Assembly [69]	10	9	19
	Bin picking [59, 70, 71]	26	17	43
	Defect scanning [30]	10	0	10
Surgical	Knot Tying [5]	19	11	30
	Debridement [26]	11	5	16
Open Datasets	Agibot [72]	10	4	14
	DROID [73]	52	17	69
	Open X-Embodiment [21]	45	22	67
Total		**318**	**156**	**474**

Fig. 4: **RoboVista Gallery**. Representative visual observations and tasks across the six domains, including agriculture, autonomous driving, domestic, industrial, surgical, and open robot datasets. The table summarizes task types and the distribution of perception- and planning-centric questions.

TABLE I: **Zero-Shot Performance of Multimodal Foundation Models on RQA (%).**

Model	All	Agriculture	Driving	Home	Industry	Surgery	Open Datasets
Baselines							
Random	20.0	20.0	20.0	20.0	20.0	20.0	20.0
Qwen 3-8B (Text-Only)	25.1	27.4	30.0	30.8	22.2	26.1	24.0
Closed-Source Models							
GPT-4o	49.6	**50.0**	50.0	59.2	32.5	67.4	53.5
GPT-5	48.1	38.7	**55.0**	46.1	35.7	63.0	**58.3**
Gemini 2.5 Pro	**56.5**	48.4	50.0	**63.2**	**48.4**	**76.1**	**58.3**
Open-Source Models							
Qwen 2.5 VL-3B	39.9	33.9	40.0	43.4	27.8	50.0	47.9
Qwen 2.5 VL-7B	43.7	37.1	45.0	47.4	32.5	52.2	51.4
Qwen 2.5 VL-32B	43.9	**48.4**	50.0	46.1	31.0	58.7	46.5
Qwen 2.5 VL-72B	44.3	43.5	35.0	40.8	31.7	**69.6**	50.7
Robo2VLM-ER	42.6	35.5	40.0	43.4	32.5	54.3	50.7
RoboBrain 2.5-8B	45.8	37.1	55.0	51.3	36.5	56.5	50.0
RoboBrain 2.5-8B-NV	47.0	27.4	40.0	**56.6**	38.9	65.2	52.8
Qwen 3 VL-4B	44.3	38.7	45.0	42.1	34.9	54.3	52.8
Qwen 3 VL-8B	46.8	38.7	45.0	**56.6**	32.5	58.7	54.2
Qwen 3 VL-32B	49.2	**48.4**	55.0	48.7	35.7	65.2	55.6
Qwen 3-235B-A22B	**51.3**	46.8	**60.0**	53.9	**37.3**	**69.6**	**56.9**

Differences in Application Domains Across domains, performance varies substantially depending on the underlying visual properties and question difficulties. Domestic robotics consistently yields higher accuracies for most models, as these scenes are typically well-lit, human-scale, and visually structured. In contrast, agricultural robotics remains one of the most challenging domains. Agricultural scenes often involve fine-grained plant morphology, severe self-occlusion, and large intra-class appearance variation across growth stages, species, and environmental conditions. Questions in this domain frequently require reasoning about deformable, partially observable structures (e.g., leaves, stems, weeds) and subtle geometric cues that are weakly represented in existing VLM pretraining corpora. These factors lead to consistently lower performance, even for large-scale and robotics-specialized models.

TABLE II: **Chain-of-Thought Effect on Perception and Planning.** Comparing zero-shot baseline vs CoT prompting. Green = improvement, Red = degradation. CoT consistently reduces scene understanding accuracy due to over-thinking but performs better with planning related tasks.

	Overall			Perception			Planning		
Model	Base	CoT	Δ	Base	CoT	Δ	Base	CoT	Δ
Qwen 2.5 VL-3B	39.9	29.5	-10.3	41.2	29.2	-11.9	37.2	30.1	-7.1
Qwen 2.5 VL-7B	43.7	36.5	-7.2	45.9	34.6	-11.3	39.1	40.4	+1.3
Qwen 2.5 VL-32B	43.9	37.6	-6.3	44.7	38.4	-6.3	42.3	35.9	-6.4
Qwen 2.5 VL-72B	44.3	43.0	-1.3	45.6	41.2	-4.4	41.7	46.8	+5.1
Robo2VLM-ER	42.6	37.6	-5.1	45.6	36.8	-8.8	36.5	39.1	+2.6
RoboBrain 2.5-8B	45.8	40.7	-5.1	47.2	40.3	-6.9	42.9	41.7	-1.3
RoboBrain 2.5-8B-NV	47.0	41.6	-5.5	48.1	40.9	-7.2	44.9	42.9	-1.9
Qwen 3 VL-4B	44.3	44.9	+0.6	46.9	45.6	-1.3	39.1	43.6	+4.5
Qwen 3 VL-8B	46.8	44.9	-1.9	47.2	46.9	-0.3	46.2	41.0	-5.1
Qwen 3 VL-32B	50.4	52.1	+1.7	53.1	51.6	-1.6	44.9	53.2	+8.3
GPT-5	55.5	55.7	+0.2	56.6	56.3	-0.3	53.2	54.5	+1.3

B. *Results on Chain-of-Thoughts Prompting*

Chain-of-Thought (CoT) prompting [23] is a widely used technique for eliciting explicit intermediate reasoning in large language models and has shown performance increase in some robotics tasks [10]. We evaluate CoT by appending the following instruction to the end of each question: *"Think step by step about this question, then provide your final answer."* All other prompting and decoding settings are kept identical to the zero-shot evaluation.

Effect on Scene Understanding. Table II shows the effect of CoT prompting on Scene Understanding questions, measured as the accuracy difference between CoT and zero-shot prompting. Across most models and domains, CoT consistently reduces scene understanding accuracy, with overall drops of up to 12%. The degradation is most significant in domains such as Driving and Domestic robotics, where questions require precise and direct spatial perception. This follows similar observation as an empirical study on VLM thinking [79] that enforcing explicit reasoning traces sometimes lead the models to overly attend to language semantics and can interfere with fine-grained perceptual processing.

TABLE III: **In-Context Learning Effect.** Comparing zero-shot baseline vs ICL with reasoning examples. Calibration Error (CE) measures overconfidence as a typical indicator for hallucination [80] (↓ = better).

	Accuracy (%)↑			Calibration Error (%)↓		
Model	Base	ICL	Δ	Base	ICL	Δ
Qwen 2.5 VL-3B	39.7	35.8	-3.8	46.3	51.3	+5.0
Qwen 2.5 VL-7B	43.0	39.1	-4.0	41.9	46.2	+4.4
Qwen 2.5 VL-32B	42.0	39.2	-2.8	50.8	54.5	+3.7
RoboBrain-MT	47.5	41.0	-6.5	38.7	48.4	+9.7
RoboBrain-NV	47.3	42.7	-4.6	37.2	46.0	+8.8
Qwen 3 VL-4B	43.0	39.3	-3.7	49.7	54.1	+4.4
Qwen 3 VL-8B	47.3	42.3	-5.0	47.2	52.8	+5.6
Qwen 3 VL-32B	48.3	46.6	-1.7	44.6	47.7	+3.1

Effect on Planning. In contrast to its impact on perception, CoT generally improves planning performance for several models, particularly in multi-step domains such as Surgical and Agricultural robotics. These tasks often require sequencing actions, reasoning over task constraints, or anticipating future outcomes, where explicit intermediate reasoning can be beneficial. However, completely isolating planning from perception is not possible, so balancing the right amount of thinking is important for the planning tasks with VLMs.

C. Results on In-Context Learning

In-context learning (ICL) [25] refers to conditioning a model on a small number of example question–answer pairs provided directly in the prompt, without any parameter updates. ICL has been shown to improve performance in general VQA and reasoning benchmarks by implicitly adapting model behavior to the target task distribution. We evaluate ICL by prepending a small set of representative Robot-VQA examples with explicit reasoning steps to the prompt. We provide a random example on the same domain and task, and average the accuracy across five queries. All other evaluation settings remain identical to the zero-shot baseline.

In-Context Learning Effects. Table III summarizes the effect of ICL. Even providing in-context examples of the same domain and task, ICL consistently *reduces* accuracy across all evaluated models, with drops ranging from 2.8% to 6.5%. We observe the degradation for both general-purpose and robotics-specialized VLMs. These results suggest that, unlike abstract reasoning benchmarks, providing in-context examples about spatial relationship may distract VLMs in longer contexts and lead to hallucination. However, we also recognize that larger models with stronger reasoning capabilities can be less affected by the in-context examples.

In-Context Learning can lead to hallucination. To further analyze the effect of ICL, we examine hallucination with model calibration [80]. To measure calibration, we prompt models to provide both an answer and their confidence from 0% to 100% with the setup as Wei et al [80] and Humanity's Last Exam [81]. Intuitively, a well-calibrated model should express high confidence only when it is likely to be correct; systematic over-confidence indicates a tendency to hallucinate plausible but incorrect answers. As shown in Table III, ICL consistently *increases* calibration error across all models, with absolute CE increases of up to 9.7%. This trend indicates that ICL encourages models to produce more confident but less reliable predictions, which could amplify hallucinated reasoning when visual evidence is ambiguous or weakly grounded.

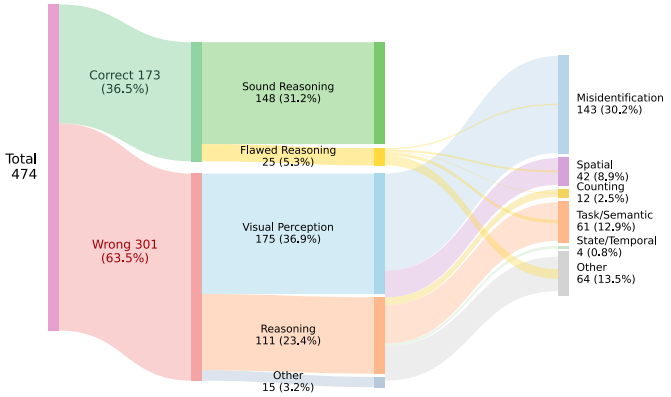


Fig. 5: **Failure Analysis of Qwen2.5-VL 7B**. We analyze the reasoning chain of Qwen2.5-VL 7B and categorize failures in misidentification and reasoning.

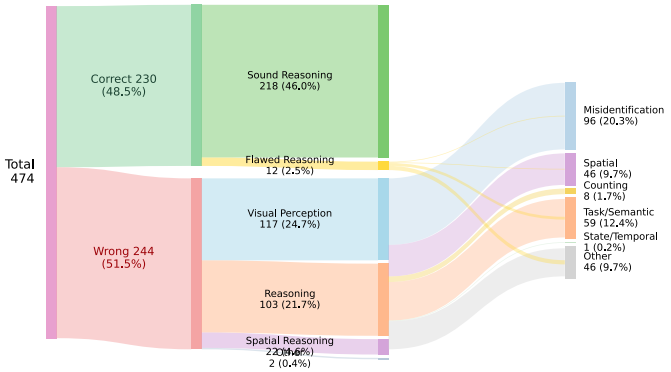
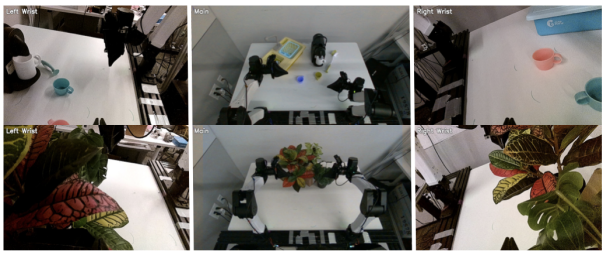


Fig. 6: **Failure Analysis of Qwen3-VL 235B**. We analyze the reasoning chain of Qwen3-VL 235B and categorize failures in misidentification and reasoning.

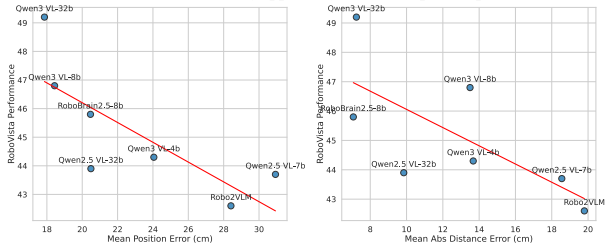
D. Results on Failure Analysis

We conduct failure analysis using model-generated reasoning chains that separates visual perception errors from higher-level reasoning failures. For both correct and incorrect predictions, we use Qwen3-VL-32B-Thinking to propose an initial failure category given groundtruth human reasoning trace and answer. Then we manually verify and correct these assignments.

Failure Analysis on Qwen2.5-VL-7B. Figure 5 shows the failure breakdown. A key observation is that the majority of failures originate from visual perception rather than logical reasoning. The most frequent failure mode is misidentification, accounting for 143 cases (30.2%), primarily involving incorrect object identity, state, or spatial location. Common examples include confusing visually similar objects, or incorrectly localizing target objects under occlusion or clutter. Spatial reasoning errors form the second largest category, encompassing failures in relative positioning, depth estimation, motion inference, and reachability. Task and semantic misinterpretation reflects failures in mapping visual evidence to task intent, such



(a) Two bi-manual gripper alignment setup examples.



(b) Correlation between RoboVista performance and (left) final gripper position error and (right) mean absolute distance estimation error.

Fig. 7: **Bi-manual gripper alignment task Setup and correlation with RoboVista performance.** The VLM estimates the distance between gripper tips and plans motions based on spatial priors. Higher RoboVista scores strongly correlates to lower estimation and execution errors.

as misunderstanding which object is being manipulated or misinterpreting contact and interaction states.

Failure Analysis on Qwen3-235B-A22B . We perform the same analysis for the much larger 235B model. Scaling significantly reduces both the overall error rate and the proportion of flawed reasoning chains, with correct answers increasing from 36.5% to 48.5%. Misidentification errors drop substantially (from 30.2% to 20.3%), indicating improved visual recognition and grounding at scale. However, spatial and semantic errors persist, and remain a dominant source of failure even for the largest model. This suggests that while increased model capacity improves visual robustness and reasoning consistency, fundamental challenges in spatial understanding and geometry-aware perception are not fully resolved by scale alone.

VI. Physical Evaluation

In this evaluation, we examine how VLMs perform on RoboVista benchmark correlates with two representative robotic tasks. Specifically, we study: (1) bimanual gripper position alignment, (2) surgical knot-tying.

A. Bimanual Gripper Position Alignment

In this experiment, we study how VLMs perform on 3D spatial understanding and symbolic planning tasks.

Setup. As shown in Fig. 7a, we use a bimanual manipulation task that requires VLMs to reason about spatial relationships and generate symbolic motion plans. Given visual observation from the main camera and two wrist cameras of the robot, the VLM is asked to (1) estimate the distance between the left and right gripper tips, and (2) generate a sequence of symbolic motion commands with associated numerical distances that move the right gripper to align with and touch the left gripper. We integrate VLM as following: the visual input consists of images from the left wrist camera, right wrist camera, and a top-down main camera view, as illustrated in Fig. 7a. The language query specifies the task objective and defines a robot-centric coordinate system, where forward/backward correspond to the x-axis, left/right to the y-axis, and up/down to the z-axis. The model is required to output a motion plan as a sequence of symbolic directional commands paired with metric distances, following a fixed format, e.g., $\{(\texttt{left}, 10\,\text{cm}), (\texttt{up}, 5\,\text{cm})\}$. We evaluate the robot under five distinct geometrical priors: (1) kitchen objects, (2) small industrial parts, (3) no background object, (4) checkerboard (5) plants. For each prior, we set 5 configurations with the left and right arms randomly positioned to cover the workspace and to span inter-gripper distances ranging from 10 to 40,cm. Ground-truth distances are computed using the robot's joint encoders and forward kinematics.

Results. We evaluate multiple VLMs on the bi-manual alignment task and report results averaged across all robot configurations (Fig. 7b). Across both distance estimation error and the final position error after executing the predicted motion plan, we observe a strong negative correlation between benchmark performance and physical error: models with higher RoboVista scores consistently achieve lower distance estimation errors and smaller final alignment errors after execution. In particular, the position error exhibits a strong negative correlation with RoboVista performance (Pearson $r = -0.78$; Spearman $\rho = -0.93$), and distance error shows a moderate negative correlation (Pearson $r = -0.70$; Spearman $\rho = -0.75$). This indicates a strong monotonic relationship between benchmark accuracy and real-world task performance.

B. Surgical Knot-Tying using dVRK

To evaluate the relationship between benchmark performance and real-world task success, we perform integrated and closed-loop surgical knot tying using the dVRK as described in Section III-D. This experiment evaluates whether performance measured by RoboVista in a surgical domain correlates with a model's ability to support long-horizon, physically grounded task execution.

Setup. We consider a shared-autonomy setting for surgical knot tying, motivated by scenarios in which a vision-language model assists junior operators during task execution. The physical setup is simulated to allow controlled evaluation while preserving the visual and procedural structure of real knot-tying workflows. Human experts with surgical robotics experience design a sequence of VQA questions with RQA framework that correspond to critical decision points in the knot-tying process, such as tool positioning, loop formation, tension management, and error detection as shown in Fig. 3. During execution, the VLM is queried sequentially to provide guidance at each stage, and the human operator proceeds based on the model's responses unless human intervention is required.

Results. We evaluate multiple vision-language models by asking the designed questions in sequence and measuring how far the knot-tying task can progress before intervention is necessary. Results are summarized in Table IV. We quantify execution progress with at most 3 human interventions. Across all models, higher RoboVista-surgical scores are associated with greater shared autonomy task progress. Models with stronger

TABLE IV: **Surgical knot-tying performance under shared autonomy.** *RoboVista Surgery* denotes benchmark accuracy on surgical Robot-VQA tasks. *Progress w/o Intervention* is the furthest knot-tying stage completed without intervention), and *Progress w/Intervention* is the furthest stage completed with up to three interventions.

Model	RoboVista-Surgery	Progress w/o Intervention	Progress w/ Intervention
Qwen-2.5 32B	58.7	2/16	9/16
ChatGPT-5.0	63.0	2/16	9/16
Gemini 2.5 Pro	**76.1**	2/16	**15/16**

benchmark performance consistently complete more stages of the knot-tying procedure before requiring assistance. All models exhibit a common failure mode at the second question that requires pointing, where motion execution fails despite correct high-level reasoning. The visualization is illustrated in the second image of the first row in Fig. 3. The failure suggests that even minor execution errors can be amplified across consecutive decisions and the overall physical task progress is correlated with RoboVista-surgical benchmark performance.

VII. Conclusion

We presented Robot Question Answering (RQA), a modular framework that translates diverse robot application decision points into a unified, robot-centric VQA interface, and instantiated it as RoboVista, a curated benchmark spanning six real-world robotic domains and 39 task types. Through comprehensive evaluation, we show that state-of-the-art VLMs exhibit substantial and persistent performance gaps across domains and reasoning stages.

Limitations. Constructing high-quality Robot-VQA instances requires significant effort from human domain experts to ensure visual grounding, task relevance, and unambiguous answers. While parts of the pipeline can be automated, in future work we will explore automated question generation and verification, and extend RoboVista to additional robot tasks, embodiments, and interactions.

References

[1] E. Solowjow, I. Ugalde, Y. Shahapurkar, J. Aparicio, J. Mahler, V. Satish, K. Goldberg, and H. Claussen, "Industrial robot grasping with deep learning using a programmable logic controller (plc)," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 97–103.

[2] S. Adebola, T. Sadjadpour, K. El-Refai, W. Panitch, Z. Ma, R. Lin, T. Qiu, S. Ganti, C. Le, J. Drake, et al., "Automating deformable gasket assembly," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2024, pp. 4146–4153.

[3] S. Xie, K. Goldberg, and D. Song, "Energy efficient planning for repetitive heterogeneous tasks in precision agriculture," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 7139–7145.

[4] S. Adebola, R. Parikh, M. Presten, S. Sharma, S. Aeron, A. Rao, S. Mukherjee, T. Qu, C. Wistrom, E. Solowjow, et al., "Can machines garden? systematically comparing the alphagarden vs. professional horticulturalists," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 779–11 785.

[5] Z. Chen, K. Hari, T. Dasari, K. Shieh, R. Jain, D. M. Fer, G. Guthart, and K. Goldberg, "Surgical d-knot: Augmented dexterity for tying double knots by monitoring optical flow in monocular attention windows," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2025, pp. 2148–2155.

[6] K. Hari, Z. Chen, H. Kim, and K. Goldberg, "Stitch 2.0: Extending augmented suturing with ekf needle estimation and thread management," *IEEE Robotics and Automation Letters*, vol. 10, no. 12, pp. 12 700–12 707, 2025.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, 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.

[8] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, et al., "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

[9] C. Si, Y. Zhang, R. Li, Z. Yang, R. Liu, and D. Yang, "Design2Code: Benchmarking multimodal code generation for automated front-end engineering," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds., Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 3956–3974.

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

[11] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[12] C. Ning, K. Fang, and W.-C. Ma, "Prompting with the future: Open-world model predictive control with interactive digital twins," in *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[14] B. S. Kim, J. Kim, D. Lee, and B. Jang, "Visual question answering: A survey of methods, datasets, evaluation, and challenges," *ACM Comput. Surv.*, vol. 57, no. 10, May 2025.

[15] 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.

[16] 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.

[17] X. Yue, Y. Ni, T. Zheng, K. Zhang, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al., "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9556–9567.

[18] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, et al., "MMMU-pro: A more robust multi-discipline multimodal understanding benchmark," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 15 134–15 186.

[19] K. Chen, S. Xie, Z. Ma, P. R. Sanketi, and K. Goldberg, "Robo2vlm: Improving visual question answering using large-

scale robot manipulation data," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[20] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An, et al., "Robobrain: A unified brain model for robotic manipulation from abstract to concrete," *CVPR*, 2025.

[21] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, et al., *Open X-Embodiment: Robotic learning datasets and RT-X models*, https://arxiv.org/abs/2310.08864, 2023.

[22] J. Yu, T. Sadjadpour, A. O'Neill, M. Khfifi, L. Y. Chen, R. Cheng, M. Z. Irshad, A. Balakrishna, T. Kollar, and K. Goldberg, "Manip: A modular architecture for integrating interactive perception for robot manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2024, pp. 1283–1289.

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[24] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, et al., "A survey on in-context learning," in *Proceedings of the 2024 conference on empirical methods in natural language processing*, 2024, pp. 1107–1128.

[25] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[26] Z. Chen, K. Hari, and K. Goldberg, *Surgical debridement dataset*, https://github.com/open-h-embodiment/data-collection, GitHub repository, 2025.

[27] S. Xie, C. Hu, M. Bagavathiannan, and D. Song, "Toward robotic weed control: Detection of nutsedge weed in bermudagrass turf using inaccurate and insufficient training data," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7365–7372, 2021.

[28] S. Xie, C. Hu, D. Wang, J. Johnson, M. Bagavathiannan, and D. Song, "Coupled active perception and manipulation planning for a mobile manipulator in precision agriculture applications," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[29] S. Adebola, C. M. Kim, J. Kerr, S. Xie, P. Akella, J. L. S. Rincon, E. Solowjow, and K. Goldberg, "Botany-bot: Digital twin monitoring of occluded and underleaf plant structures with gaussian splats," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2025, pp. 1839–1846.

[30] T. Qiu, Z. Ma, K. El-Refai, H. Shah, C. M. Kim, J. Kerr, and K. Goldberg, "Omni-scan: Creating visually-accurate digital twin object models using a bimanual robot with handover and gaussian splat merging," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2025, pp. 18 782–18 789.

[31] A. Adler, A. Ahmad, Y. Qiu, S. Wang, W. C. Agboh, E. Llontop, T. Qiu, J. Ichnowski, T. Kollar, R. Cheng, et al., "The teenager's problem: Efficient garment decluttering as probabilistic set cover," *arXiv preprint arXiv:2310.16951*, 2023.

[32] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "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.

[33] 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].

[34] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al., "Openvla: An open-source vision-language-action model," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[35] P. I. Team, "π_0: A generalist robot policy," *arXiv preprint arXiv:2410.24164*, 2024.

[36] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," *Autonomous Robots*, vol. 47, pp. 1345–1365, 2023.

[37] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., "PaLM-E: An embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023, pp. 8469–8488.

[38] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," *arXiv preprint arXiv:2409.01652*, 2024.

[39] K. Fang, F. Liu, P. Abbeel, and S. Levine, "Moka: Open-world robotic manipulation through mark-based visual prompting," *Robotics: Science and Systems XX*, 2024.

[40] G. Tang, S. Rajkumar, Y. Zhou, H. R. Walke, S. Levine, and K. Fang, "Kalie: Fine-tuning vision-language models for open-world manipulation without robot data," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 9507–9515.

[41] J. Shi, R. Yang, K. Chao, B. S. Wan, Y. S. Shao, J. Lei, J. Qian, L. Le, P. Chaudhari, K. Daniilidis, et al., "Maestro: Orchestrating robotics modules with vision-language models for zero-shot generalist robots," in *NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI*, 2025.

[42] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2263–2279.

[43] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, *Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts*, 2024. arXiv: 2310.02255 [cs.CV].

[44] 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.

[45] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, S. Gould, and A. van den Hengel, "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.

[46] E. Zhao, V. Raval, H. Zhang, J. Mao, Z. Shangguan, S. Nikolaidis, Y. Wang, and D. Seita, "Manipbench: Benchmarking vision-language models for low-level robot manipulation," *Conference on Robot Learning*, 2025.

[47] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang, et al., "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," in *Robotics: Science and Systems (RSS) 2025*, Robotics: Science and Systems Foundation, 2025.

[48] Y. Yuan, M. He, M. A. Shahid, Z. Li, J. Huang, and L. Zhang, "TurnaboutLLM: A deductive reasoning benchmark from detective games," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds., Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 1951–1965.

[49] Z. R. Sprague, X. Ye, K. Bostrom, S. Chaudhuri, and G. Durrett, "MuSR: Testing the limits of chain-of-thought with multistep

soft reasoning," in *The Twelfth International Conference on Learning Representations*, 2024.

[50] E. Glazer, E. Erdil, T. Besiroglu, D. Chicharro, E. Chen, A. Gunning, C. F. Olsson, J.-S. Denain, A. Ho, E. de Oliveira Santos, et al., *Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai*, 2025. arXiv: 2411.04872 [cs.AI].

[51] B. Siciliano and O. Khatib, *Handbook of Robotics*. Springer, 2016, Part C: Sensing and Perception, pp. 525-894.

[52] S. M. LaValle, *Planning Algorithms*. Cambridge, UK: Cambridge University Press, 2006.

[53] K. M. Lynch and F. C. Park, *Modern Robotics: Mechanics, Planning, and Control*, 1st. USA: Cambridge University Press, 2017.

[54] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 1987.

[55] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 39–48.

[56] T. Lozano-Pérez, "Spatial planning: A configuration space approach," *IEEE Transactions on Computers*, no. 2, pp. 108–120, 1983.

[57] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.

[58] N. Ratliff, M. Zucker, J. A. Bagnell, and S. Srinivasa, "Chomp: Gradient optimization techniques for efficient motion planning," in *2009 IEEE International Conference on Robotics and Automation*, IEEE, 2009, pp. 489–494.

[59] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, eaau4984, 2019.

[60] M. Presten, R. Parikh, S. Aeron, S. Mukherjee, S. Adebola, S. Sharma, M. Theis, W. Teitelbaum, and K. Goldberg, "Automated pruning of polyculture plants," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2022, pp. 242–249.

[61] C. Hu, S. Xie, D. Song, J. A. Thomasson, R. G. H. IV, and M. Bagavathiannan, "Algorithm and system development for robotic micro-volume herbicide spray towards precision weed management," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 633–11 640, 2022.

[62] D. Wang, C. Hu, S. Xie, J. Johnson, H. Ji, Y. Jiang, M. Bagavathiannan, and D. Song, "Toward precise robotic weed flaming using a mobile manipulator with a flamethrower," *arXiv preprint arXiv:2407.04929*, 2024.

[63] GM/SAE AutoDrive Challenge I (Year 4), *Gm/sae autodrive challenge i (year 4) yearbook*, https://online.flippingbook.com/view/101355423/, Online flippingbook yearbook for the Year 4 competition of the AutoDrive Challenge, 2021.

[64] Y. Huang, C. Agia, J. Wu, T. Hermans, and J. Bohg, "Points2plans: From point clouds to long-horizon plans with composable relational dynamics," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 1208–1216.

[65] Y. Huang, N. Alvina, M. Devendran Shanthi, and T. Hermans, "Fail2progress: Learning from real-world robot failures with stein variational inference," in *Conference on Robot Learning (CoRL), 2025*, 2025.

[66] V. Viswanath, K. Shivakumar, M. Parulekar, J. Ajmera, J. Kerr, J. Ichnowski, R. Cheng, T. Kollar, and K. Goldberg, "Handloom: Learned tracing of one-dimensional objects for inspection and manipulation," in *Conference on Robot Learning*, PMLR, 2023, pp. 341–357.

[67] L. Y. Chen, B. Shi, R. Lin, D. Seita, A. Ahmad, R. Cheng, T. Kollar, D. Held, and K. Goldberg, "Bagging by learning to singulate layers using interactive perception," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 3176–3183.

[68] L. Y. Chen, B. Shi, D. Seita, R. Cheng, T. Kollar, D. Held, and K. Goldberg, "Autobag: Learning to open plastic bags and insert objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3918–3925.

[69] A. Goldberg, K. Kondap, T. Qiu, Z. Ma, L. Fu, J. Kerr, H. Huang, K. Chen, K. Fang, and K. Goldberg, "Blox-net: Generative design-for-robot-assembly using vlm supervision, physics simulation, and a robot with reset," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 15 493–15 500.

[70] Z. Tam, K. Dharmarajan, T. Qiu, Y. Avigal, J. Ichnowski, and K. Goldberg, "Bomp: Bin-optimized motion planning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2024, pp. 11 056–11 063.

[71] W. C. Agboh, S. Sharma, K. Srinivas, M. Parulekar, G. Datta, T. Qiu, J. Ichnowski, E. Solowjow, M. Dogar, and K. Goldberg, "Learning to efficiently plan robust frictional multi-object grasps," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 10 660–10 667.

[72] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, X. He, X. Huang, et al., "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2025.

[73] A. Sharma, V. Sundaresan, Y. Zhu, P. Shah, K. Liu, M. Laskin, J. Tompson, A. Wahid, Y. Chebotar, and K. Hausman, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2310.01894*, 2023.

[74] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio, "An open-source research kit for the da vinci® surgical system," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 6434–6439.

[75] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al., "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[76] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[77] H. Tan, E. Zhou, Z. Li, Y. Xu, Y. Ji, X. Chen, C. Chi, P. Wang, H. Jia, Y. Ao, et al., "Robobrain 2.5: Depth in sight, time in mind," *arXiv preprint arXiv:2601.14352*, 2026.

[78] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. 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.

[79] M. Li, J. Zhong, S. Zhao, Y. Lai, H. Zhang, W. B. Zhu, and K. Zhang, "To think or not to think: A study of thinking in rule-based visual reinforcement fine-tuning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[80] J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus, "Measuring short-form factuality in large language models," *arXiv preprint arXiv:2411.04368*, 2024.

[81] L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, et al., "Humanity's last exam," *arXiv preprint arXiv:2501.14249*, 2025.