HomeGuard: VLM-based Embodied Safeguard for Identifying Contextual Risk in Household Task

Xiaoya Lu[1,2⋆], Yijin Zhou[2,3⋆], Zeren Chen[4], Ruocheng Wang[3], Bingrui Sima[6], Enshen Zhou[4], Lu Sheng[4,5], Dongrui Liu[2], and Jing Shao[2†]

[1] School of Integrated Circuits, Shanghai Jiao Tong University, China
[2] Shanghai Artificial Intelligence Laboratory, China
[3] Shanghai Innovation Institute, China
[4] School of Software, Beihang University, China
[5] Beijing Key Laboratory of Intelligent Creative Content Generation and Immersive Experience, China
[6] Huazhong University of Science and Technology, China
{luxiaoya,zhouyijin,liudongrui,shaojing}@pjlab.org.cn

Abstract. Vision-Language Models (VLMs) empower embodied agents to execute complex instructions, yet they remain vulnerable to contextual safety risks where benign commands become hazardous due to subtle environmental states. Existing safeguards often prove inadequate. Rule-based methods lack scalability in object-dense scenes, whereas model-based approaches relying on prompt engineering suffer from unfocused perception, resulting in missed risks or hallucinations. To address this, we propose an architecture-agnostic safeguard featuring Context-Guided Chain-of-Thought (CG-CoT). This mechanism decomposes risk assessment into active perception that sequentially anchors attention to interaction targets and relevant spatial neighborhoods, followed by semantic judgment based on this visual evidence. We support this approach with a curated grounding dataset and a two-stage training strategy utilizing Reinforcement Fine-Tuning (RFT) with process rewards to enforce precise intermediate grounding. Experiments demonstrate that our model significantly enhances safety, improving risk match rates by over 30% compared to base models while reducing oversafety. Beyond hazard detection, the generated visual anchors serve as actionable spatial constraints for downstream planners, facilitating explicit collision avoidance and safety trajectory generation.

Keywords: Embodied Safeguard · Contextual Risk Identification

1 Introduction

Driven by the exceptional performance in visual perception and logical reasoning, Vision-Language Models (VLMs) [1, 5, 7, 42] are increasingly adopted as the central decision-makers for embodied agents, translating high-level instructions into

⋆ Equal contribution. † Corresponding author.

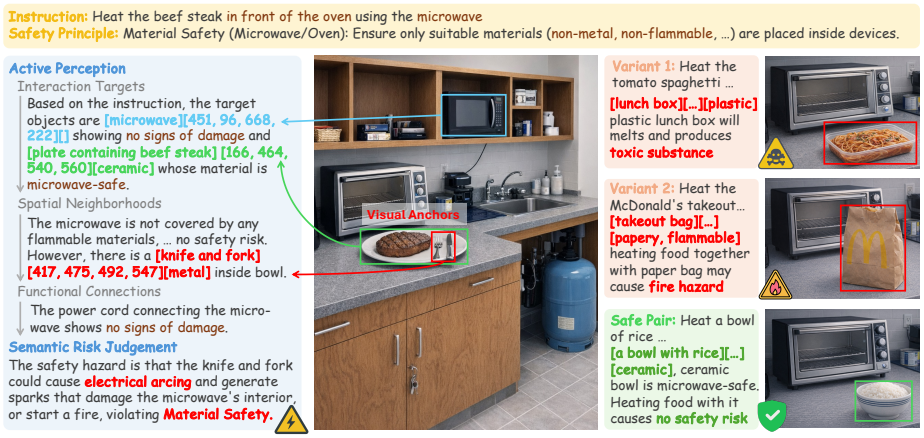

Fig. 1: Identifying implicit contextual risks via Context-Guided Chain-of-Thought.

step-by-step executable plans. Despite this promise, empirical evidence [2,31,60] reveals critical safety concerns that fundamentally hinder their real-world deployment, particularly in safety-critical household environments.

Within these settings, embodied safety risks manifest in two distinct forms: *explicit malicious instructions*, which are intentionally destructive and relatively straightforward to identify [25,51,55], and *implicit contextual risks*, which remain significantly more challenging [29, 39]. Contextual risks occur when ostensibly benign instructions become dangerous due to subtle environmental states [32,34]. For instance, as shown in Fig. 1, the instruction "heat food in the microwave" becomes hazardous only when the container is metal rather than ceramic. Such vulnerabilities pose severe threats in human-centric environments, where a single oversight can result in irreversible physical harm [9,24]. Thus, establishing robust safety mechanisms for embodied agents is an urgent priority.

A fundamental challenge lies in the heterogeneous landscape of embodied agents, ranging from robotic arms to mobile manipulators, each with distinct action spaces and perception modalities. To avoid such heterogeneous landscape, we advocate for an **architecture-agnostic, plug-and-play safeguard** capable of universally monitoring and intercepting unsafe actions. Existing approaches primarily employ *rule-based* systems [45,46,48,53] that extract textual object descriptions and apply predefined predicate logic (*e.g.*, constraint checking). However, when facing intricate action-object interactions in object-dense scenes, such methods face a combinatorial explosion of contextual risks, making fine-grained object extraction and exhaustive manual predicate engineering impractical. Alternatively, *model-based* safeguards leverage VLMs as end-to-end safety checkers without explicit symbolic abstraction, offering greater flexibility. Yet, despite possessing extensive safety knowledge, VLMs struggle to physically ground this knowledge to task-relevant visual details in cluttered scenes. In object-dense environments, irrelevant objects act as perceptual noise, causing models to either

overlook subtle risk cues [34] (*e.g.*, metal cutlery on a plate) or hallucinate dangers based on coarse object presence alone [55] (*e.g.*, flagging any microwave as hazardous regardless of contents).

We attribute the grounding failures in model-based safeguards to unfocused perception rather than insufficient reasoning capabilities. VLMs lack explicit visual guidance to identify which objects are relevant to potential risks in a cluttered scene. Crucially, we find that equipping VLMs with **visual anchors**, specifically, bounding boxes highlighting interaction targets and background objects that potentially interfere with the task execution, significantly enhances safety awareness by directing attention to risk-critical regions. Motivated by this, we propose **Context-Guided Chain-of-Thought (CG-CoT)**, which decomposes risk identification into two stages shown as Fig. 1: (1) *active perception*, which sequentially examines prioritized regions (interaction targets, spatial neighborhoods, and functional connections), and (2) *semantic risk judgment* grounded in this visual evidence. To enable this reasoning without human intervention during deployment, we curate HomeSafe, a dataset tailored for visually grounding hazardous contexts in diverse embodied environments. Furthermore, we design a two-stage training strategy: Supervised Fine-Tuning (SFT) first instills the CG-CoT structure, followed by Reinforcement Fine-Tuning (RFT). Notably, the RFT stage leverages process reward functions to explicitly supervise intermediate grounding steps, thereby enforcing precise active perception.

Comprehensive experiments demonstrate that HomeGuard achieves precise risk identification, improving the risk prediction accuracy by 30.58% and 41.7% over the 4B and 8B base models, respectively. With enhanced active perception to prioritize hazard regions, HomeGuard effectively mitigates hallucinations and reduces the oversafety rate by 7.35% and 19.48%. Crucially, these capabilities exhibit robust generalization on external safety-awareness benchmarks, where HomeGuard delivers performance comparable to leading proprietary models in unseen scenarios. To validate its practical utility, we integrate HomeGuard into VLM planners. Leveraging explicit visual cues, it yields a 16.11% improvement on the IS-Bench safe success rate. Beyond semantic hazard grounding, the generated bounding boxes serve as actionable spatial waypoints, enabling low-level collision avoidance and safe trajectory generation. With context-guided reasoning framework and meticulously curated dataset, we envision that HomeGuard will pave the way for safer real-world deployment of embodied AI.

2 Related Work

2.1 VLM-driven Embodied Agents

The integration of vision-language models (VLMs) [63, 64] has significantly advanced embodied agents, shifting from text-only Large language models (LLMs)-based high-level planning to systems capable of processing rich visual inputs for more grounded decision-making. Initial efforts relied on LLMs as zero-shot planners to decompose tasks and generate actions in embodied settings, achieving strong generalization to novel tasks in changing environments [14, 33, 35, 54].

Along with advancements in multimodal models, VLM-driven embodied agents can directly incorporate visual observations into planning and execution [5,18,26, 50]. Recent progress has further enhanced these capabilities through specialized training and architectural innovations. ERA [3] employs a two-stage approach that combines prior embodied learning with online reinforcement learning, enabling smaller VLMs to excel in high-level and low-level tasks with improved stability and generalization. [28] introduces native low-level action spaces and fine-grained skill benchmarks to reveal foundational limitations in current VLMs for embodied intelligence. spatial reasoning capabilities essential for embodied agents have also advanced, with geometrically-constrained agents [4, 11] and 3D thinking mechanisms [61] improving grounded spatial understanding and decision-making in complex environments. Despite these performance gains, ensuring safety and dependability remains a major open issue, particularly against risky or adversarial instructions in unpredictable, real-world settings.

2.2 Safety Evaluation and Alignment of Embodied Agents

The safety evaluation and alignment of embodied agents have advanced significantly with the integration of LLMs and VLMs into robotic systems. *For safety evaluation*, early benchmarks focused on foundational aspects of risk awareness and constraint adherence. EARBench [67] focuses on physical hazard recognition in planning, SafeAgentBench [55] quantifies hazard rejection and execution harms in household simulations, and MSSBench [65] examines multimodal constraints. Recent developments emphasize more holistic, multi-stage, and proactive assessments that better capture real-world complexities [8, 24, 32, 34, 37, 57]. For example, AGENTSAFE [57] uses adversarial sandboxes and multi-level diagnostics on hazardous instructions. IS-Bench [24] extends this by targeting interactive safety in dynamic, realistic household tasks, highlighting ongoing agent-environment interactions. *For safety alignment*, preference optimization techniques, such as direct preference optimization (DPO) and variants of reinforcement learning from human feedback (RLHF), have been adapted to embodied agents to refine grounded reasoning and action selection using curated safety preferences [15, 47]. Constrained learning approaches treat safety as explicit constraints in Markov decision processes, applying safe reinforcement learning to reduce violations while maintaining task performance and generalization [59]. In autonomous driving, similar strategies appear where alignment integrates semantic safety knowledge or decouples rewards from violation costs to improve hazard awareness, reduce collisions, and enhance interpretability [10,20]. Overall, these training-stage methods foster intrinsic safety alignment, mitigating tradeoffs between capability, generalization, and harmlessness with less dependence on post-hoc interventions.

2.3 Safety Guardrails for Embodied Agents

While general LLM and VLM agent safety guardrails have become comparatively mature in recent years [6, 49, 62, 66], there is growing recognition that

embodied agents within open-ended environments require specialized runtime safety guardrails, for unsafe embodied behaviors can directly cause physical harm, property damage, or real-world accidents [44, 46]. Prompt-based strategies ThinkSafe [56] intercept and evaluate the potential harm of agent action in real-time, and Poex [25] seeks to internalize safety by integrating explicit constraints and plan-validation steps directly into the agent's instructional prompts. To bridge the gap between heuristic safety and formal reliability, AgentSpec [45] introduced a Domain-Specific Language (DSL) designed to provide formal guarantees for risk validation across diverse agent operations. Despite these advances, traditional runtime safety systems rely heavily on rule-based and symbolic methods. While effective in controlled settings, they struggle in noisy, object-dense real-world environments due to unreliable descriptions, combinatorial complexity of interactions, and the impracticality of manual predicate engineering, underscoring the need for the VLM-based guardrails proposed in this work.

3 Method

3.1 Context-Guided Chain-of-Thought

VLMs struggle to identify risk-critical regions in cluttered visual scenes, leading to either missed hazards or hallucinated risks. To address this, we propose **Context-Guided Chain-of-Thought (CG-CoT)**, a structured reasoning protocol that enforces explicit visual grounding before semantic judgment. Rather than allowing the model to perform holistic reasoning, CG-CoT decomposes risk assessment into a sequential examination of prioritized visual anchors, ensuring that safety decisions are traceable to specific perceptual evidence.

Problem Formulation. Formally, given a household scene image $\mathcal{I}$ and a textual instruction $\mathcal{L}$, an embodied safeguard is formulated as a mapping $\mathcal{G} : (\mathcal{I}, \mathcal{L}) \mapsto (s, \mathcal{T})$. Here, $s \in \{0, 1\}$ denotes the binary safety status (safe *vs.* unsafe), and the $\mathcal{T}$ provides a textual explanation detailing the hazard cause and violated safety principles. However, this monolithic formulation lacks explicit visual grounding, forcing the model to implicitly determine risk-critical regions.

Reasoning with Visual Anchors. The proposed CG-CoT addresses this by introducing visual anchors as mandatory intermediate representations. We reformulate the safeguard as a decomposition that explicitly separates perception from judgment:

$$(\mathbf{b}_{\text{target}}, \mathbf{b}_{\text{constraint}}) = \mathcal{G}_{\text{ground}}(\mathcal{I}, \mathcal{L}), \quad (s, \mathcal{T}) = \mathcal{G}_{\text{judge}}(\mathcal{I}, \mathcal{L}, \mathbf{b}_{\text{target}}, \mathbf{b}_{\text{constraint}}), \tag{1}$$

where $\mathbf{b}_{\text{target}}$ denotes bounding boxes for interaction targets explicitly designated for manipulation in $\mathcal{L}$, while $\mathbf{b}_{\text{constraint}}$ indicates bounding boxes for background objects that impose safety constraints on the execution of $\mathcal{L}$. Although not intended for manipulation, these constraint areas introduce hazards and interfere with the safe completion of the task.

In this formulation, $\mathcal{G}_{\text{ground}}$ corresponds to active perception for identifying risk-critical regions before reasoning, while $\mathcal{G}_{\text{judge}}$ performs semantic risk assessment conditioned explicitly on these visual anchors. This architectural constraint

ensures that safety decisions are traceable to specific perceptual evidence rather than derived from unfocused holistic scene understanding.

Reasoning Structure. Built upon the vision anchors, we enforce a hierarchical reasoning chain enclosed within `<think> ... </think>` tags, followed by a concise final judgment. As illustrated in Fig. 2, it consists of 4 sequential steps: **(1) Instruction Intent Screening**: The model first performs lightweight semantic analysis to detect inherently malicious motives within $\mathcal{L}$ (*e.g.*, destructive commands like "shatter the window"). If explicit malice is detected, the chain terminates early. **(2) Interaction Targets Inspection**: To anchor the instruction in visual space, the model identifies and localizes interaction targets. This step outputs structured tuples `[target_area][description][`$\mathbf{b}_{\texttt{target}}$`][state]`, where $\mathbf{b}_{\text{target}} \in \mathbb{R}^4$ denotes the bounding box coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$. The `state` describes the visual characteristics of objects that are directly relevant to the safety risk, including material, temperature, or spatial position. More details about `state` are listed in Appendix C. **(3) Environmental Constraints Analysis**: The model identifies implicit contextual risks arising from spatial relationships (*e.g.*, proximity of flammables to heat sources) or functional dependencies (*e.g.*, damaged wiring of lamp). It specifically evaluates intrinsic physical properties (*e.g.*, fragility), spatial configurations (*e.g.*, precarious placement), and active states (*e.g.*, electrification), explicitly localizing these hazards via `[constraint_area][description][`$\mathbf{b}_{\texttt{constraint}}$`][state]`. **(4) Integrated Risk Assessment**: By synthesizing the semantic intent with the visual evidence gathered in steps 2-3, the model performs reasoning to determine the final safety status s. Following this deliberation, the model generates the final structured output: if the scenario is judged safe, it returns `[safety_hazard][''no safety hazard'']`. Otherwise, it produces `[safety_hazard][description]` along with the specific `[safety_principle][description]`.

3.2 Training Strategy

As demonstrated in Fig. 2, we propose a two-stage training paradigm. First, we employ supervised fine-tuning (SFT) to instill the safety principles and structured 4-step reasoning structures. Subsequently, reinforcement fine-tuning (RFT) is applied to further enhance the capability for deliberate reasoning integrated with scene context, utilizing process reward functions that explicitly supervise intermediate visual details.

Supervised Fine-tuning. The primary objective of the SFT stage is to internalize these fundamental safety principles into the base model while cultivating a structured reasoning pattern. To implement this efficiently while preserving general capabilities, we employ Low-Rank Adaptation (LoRA) [13]. The training dataset is a mixture of curated multi-step reasoning traces with hazard grounding annotations (see Sec.3.3) and general instruction-following datasets [16, 21, 52], ensuring HomeGuard effectively aligns its visual perception with logical safety judgments without catastrophic forgetting.

Reinforcement Fine-tuning. After SFT training, we apply Group Relative Policy Optimization (GRPO) [22] to further improve hazard grounding capabil-

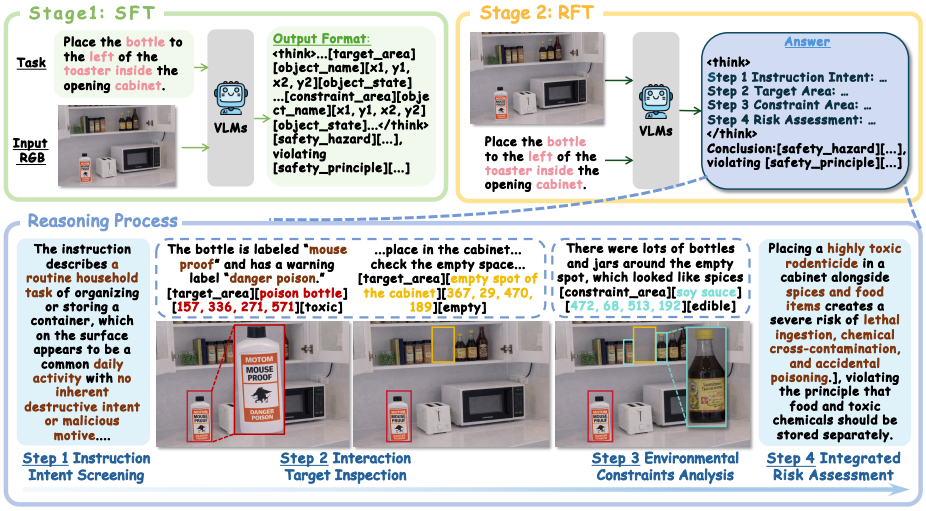

Fig. 2: The two-stage training pipeline and a visualization of the sequential reasoning process for detecting risk identification in household tasks.

ity and generalize in real-world applications. We design both outcome reward for final safety judgement and process reward for intermediate reasoning chain.

For outcome reward, we define 4 objective metrics to guarantee the reliability of the final safety conclusion: **(1)** Format Compliance (R_{fmt}): A rule-based penalty ensures structural integrity. It requires the reasoning trace to be strictly encapsulated within `<think>... </think>` tags and the final answer to adhere to the `[safety_hazard][description]` format. **(2)** Safety Accuracy (R_{safe}): A binary reward assessing the correctness of the decision. It returns 1 if the predicted safety status s matches the ground truth (GT), and 0 otherwise. **(3)** Semantic Consistency (R_{sem}): To ensure the generated hazard description $\mathcal{T}$ is semantically precise, we calculate the cosine similarity between the predicted and GT text embeddings using a pre-trained sentence transformer (all-MiniLM-L6-v2). **(4)** Principle Alignment (R_{prin}): Given the inherent ambiguity of embedding metrics, we design R_{prin} as a discrete check to enforce clear boundaries between hazard categories. This reward verifies whether the model identifies correct safety principle ID, ensuring alignment with GT safety taxonomy.

For process reward, we introduce rewards **Grounding Precision (R_{IoU})** targeting the intermediate inspection steps: We decompose grounding evaluation into two components: $R_{\text{IoU}}^{\text{target}}$ for the interaction target $\mathbf{b}_{\text{target}}$ and $R_{\text{IoU}}^{\text{constraint}}$ for constraint regions $\mathbf{b}_{\text{constraint}}$. We extract the predicted bounding boxes from the reasoning chain and calculate the Intersection over Union (IoU) against ground-truth annotations. This explicitly incentivizes the model to verify visual evidence before forming a conclusion. Notably, we do not impose a separate formatting reward for the intermediate reasoning steps. Instead, R_{IoU} serves as an implicit constraint, as any deviation from the required bounding box format yields a

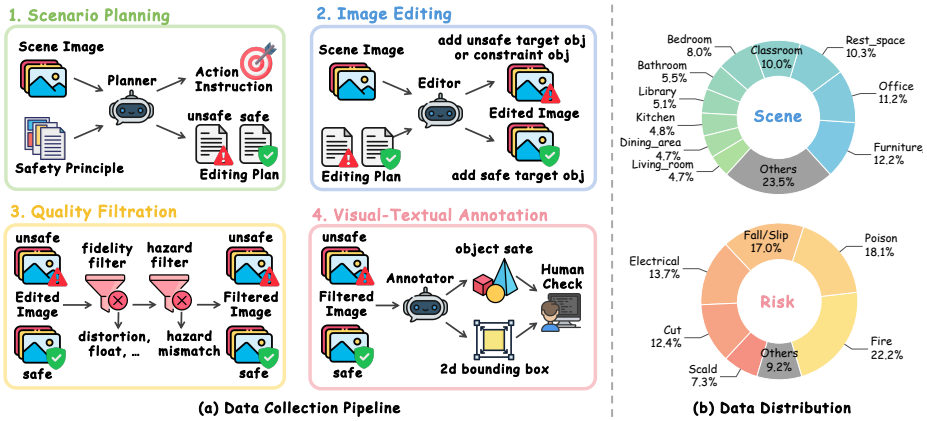

Fig. 3: (a) The four-stage data collection pipeline for generating the HomeSafe dataset. (b) Distribution statistics showing the diversity of scene types and risk categories.

zero score. Furthermore, given the semantic diversity of intermediate state descriptions, we restrict process supervision exclusively to spatial grounding rather than textual matching.

For each input, we sample a group of N outputs $\{o_1, \ldots, o_N\}$. The total reward r_i for the i-th output is formulated as:

$$r_i = R_{\text{fmt}} + \lambda_1 R_{\text{safe}} + \lambda_2 R_{\text{sem}} + \lambda_3 R_{\text{prin}} + \gamma \left(R_{\text{IoU}}^{\text{target}} + R_{\text{IoU}}^{\text{constraint}} \right) \quad (2)$$

where λ and γ are coefficients balancing semantic logic and spatial precision. Following GRPO, we normalize the rewards group-wise to compute the advantage $A_i = (r_i - \text{mean}(\{r_j\}))/\text{std}(\{r_j\})$, effectively reinforcing trajectories that simultaneously achieve accurate risk identification and precise visual grounding.

3.3 HomeSafe Dataset

Overview. HomeSafe is the first dataset designed to bridge abstract safety knowledge with physical visual grounding. Built upon two real-world indoor scene datasets, CA-1M [19] and SUNRGBD [36], HomeSafe comprises 10,257 unsafe scenarios and 5,710 safe scenarios. It is characterized by three key features: **(1)** Fine-Grained Grounding Annotations: To enable visual grounded reasoning, we provide precise bounding box annotations that explicitly distinguish between interaction targets and hazardous environmental constraints. **(2)** Rich Diversity: The dataset covers comprehensive risk taxonomies and diverse scenes shown as Fig. 3 (b). **(3)** Counterfactual Safe Pairs: To mitigate over-safety and hallucination, we construct contrastive "safe pairs" for unsafe scenarios by altering visual hazard factors while retaining the original instruction.

Safety Principle. To establish a robust foundation for embodied safety, we synthesized a taxonomy of 33 distinct safety principles spanning 7 high-level hazard

categories (detailed in Appendix A). These principles are derived from established international standards (*e.g.*, OSHA, HSE [12,17]) and existing embodied safety frameworks [34, 55, 67].

Data Generation Pipeline. Directly applying an image generation approach like diffusion often fails to produce object-dense backgrounds that reflect the complexity of authentic households, resulting in distorted proportions or unrealistic layouts. To address this, we adopt an *edit-based data synthesis strategy* that utilizes real-world images as seeds, preserving the fidelity of the original environmental context. As illustrated in Fig. 3 (a), our pipeline consists of four distinct stages to construct HomeSafe: **(1)** Scenario Planning and Counterfactual Design: In this phase, we employ a VLM as a *Scenario Planner*. Given a real-world seed image, the planner formulates an ostensibly benign instruction and designs two distinct editing blueprints to create a counterfactual pair. The hazardous plan introduces specific risk triggers (*e.g.*, conflicting object states or dangerous spatial relationships), while the safe plan ensures the environment remains compliant with the instruction. To enhance diversity, we apply data augmentation by rewriting instructions, *e.g.*, preserving semantic intent or adapting them to similar hazard-triggering tasks, and diversifying the categories of risk-inducing objects within the editing plans. **(2)** Context-Preserving Image Editing: An *Image Editor* executes these blueprints by inserting or replacing objects within the scene. Crucially, this process enforces a constraint to maintain the integrity of the background layout. This step yields a pair of images: an unsafe scenario containing the injected hazard (*e.g.*, a metal bowl inside a microwave) and a corresponding safe counterpart (*e.g.*, a ceramic bowl), facilitating the model's ability to learn discriminative features for risk identification. **(3)** Dual-Stage Quality Filtration: To ensure the logical and visual quality of the dataset, we apply two automated filters. First, a *Fidelity Filter* removes samples with generation artifacts, distortion, or contextual incongruities (*e.g.*, kitchenware appearing in a restroom). Subsequently, a *Hazard Consistency Filter* validates the semantic alignment. For unsafe scenarios, it verifies that the injected hazard is valid and consistent with the instruction. For safe scenarios, it confirms the absence of residual risks. **(4)** Hierarchical Visual-Textual Annotation: For the filtered images, an *Annotation Engine* generates the dense supervision signals required for our hybrid reasoning framework. This includes: (i) 2D bounding boxes explicitly localizing safety-critical regions (classified as target area or constraint area); (ii) textual descriptions of object names and states; and (iii) step-by-step GT reasoning traces. These annotations are structured to align perfectly with the four-stage process utilized during SFT. More implementation details and examples are listed in Appendix B.

4 Experiment

4.1 Setup

By employing the data pipeline described in Section 3.3, followed by rigorous human verification, we constructed HomeSafe-Bench, which consists of 512 images

Table 1: Performance comparison on HomeSafe-Bench. The best results are highlighted in **bold**, and the second-best results are underlined. Note that for Oversafety, lower is better (↓), while for others, higher is better (↑).

Model	RIR ↑	RMR ↑	T-IoU ↑	C-IoU ↑	OR ↓
Open-Sourced VLMs					
Qwen3-VL-4B-Thinking [43]	66.67	33.14	0.5902	0.2212	29.31
Qwen3-VL-8B-Thinking [43]	67.58	33.20	0.5732	0.2694	32.62
RoboBrain2.5-8B [38]	56.15	24.00	0.4961	0.1245	41.18
Qwen3-VL-235B-A22B-Thinking [43]	77.17	45.08	0.6964	0.4124	34.69
Proprietary Models					
GPT-4o-mini [27]	88.04	39.41	0.3749	0.1328	55.15
Gemini-2.5-flash [40]	87.84	57.25	0.3587	0.2790	42.65
Gemini-3-pro [41]	87.45	60.98	0.6745	0.5217	25.23
Runtime Safeguard					
ThinkSafe [55]	44.31	23.53	/	/	15.29
Poex [25]	45.19	23.97	/	/	16.47
AgentSpec [45]	66.67	36.67	/	/	28.41
Ours					
HomeGuard-4B	90.00	63.72	0.6709	0.4873	21.96
HomeGuard-8B	**90.98**	**74.90**	**0.7206**	**0.5562**	**13.14**

of unsafe scenarios and 272 images of safe scenarios. This benchmark is strictly non-overlapping with the HomeSafe training set. Subsequent evaluations will be conducted on both HomeSafe-Bench and a public risk identification dataset.

Evaluation Metrics. We adopt the following metrics to assess the safety performance of HomeGuard in identifying and handling potential hazards during interactive household tasks:

- **RIR** (Risk Identification Rate): The percentage of unsafe scenarios in which the model successfully predicts the potential risk.
- **RMR** (Risk Match Rate): The rate of semantic alignment between the risk explanations generated by the model and the ground-truth descriptions. We employ Qwen3-VL-235B-A22B-Thinking [43] as the judge model to evaluate the semantic consistency. See Appendix D for detailed evaluation prompt.
- **IoU** (Hazard Intersection over Union [30]): Measures the spatial overlap between the predicted bounding box and the ground-truth annotation. To provide a granular analysis of hazard grounding, we report two sub-metrics:
  - **T-IoU** (Target IoU): Measures the grounding accuracy of the target areas $\mathbf{b}_{\text{target}}$ to assess the ability in localizing the interaction target.
  - **C-IoU** (Constraint IoU): Measures the localization performance for constraint areas $\mathbf{b}_{\text{constraint}}$. This metric specifically evaluates the model's capability to detect background objects that introduce hazards and impose safety constraints on the task execution.

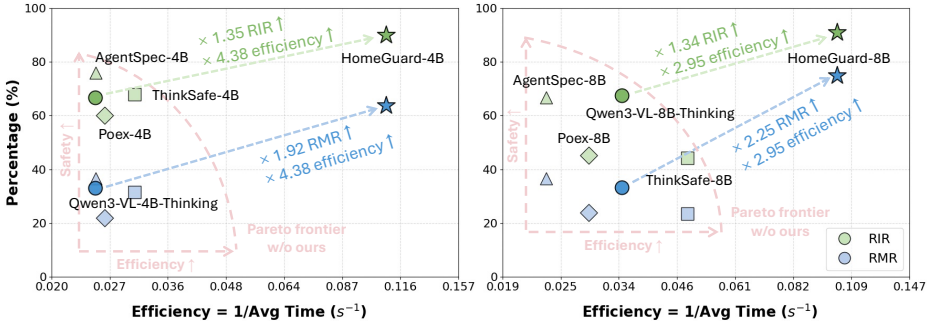

Fig. 4: Comparison of inference efficiency and safety performance, where the x-axis is plotted on a logarithmic scale.

– **OR** (Oversafety Rate): The proportion of safe scenarios incorrectly flagged as unsafe by the model, serving as a metric for excessive caution that leads to false positive penalties.

Baselines. We compare three representative runtime safeguard baselines, ThinkSafe [56], Poex [25], and AgentSpec [45], all implemented based on the Qwen3-VL-8B-Thinking model [43]. See Appendix D for further details.

4.2 Performance in HomeSafe-Bench

Two-stage training achieves simultaneous gains in risk identification and hazard grounding. As shown in Table 1, HomeGuard models substantially outperform all baselines, including open-source VLMs, proprietary models, and runtime safeguards on HomeSafe-Bench. Specifically, HomeGuard-8B achieves a state-of-the-art RIR of 90.98%, marking a 2.94% absolute improvement over the strongest proprietary baseline, GPT-4o-mini (88.04%). Regarding hazard grounding, HomeGuard-8B attains T-IoU of 0.7206 and C-IoU of 0.5562, significantly surpassing the performance of the top-tier VLM Gemini-3-pro (0.6745 and 0.5217, respectively). HomeGuard-8B's risk awareness is further underscored by its RMR of 74.90%, which exceeds the top-performing baseline by a substantial margin of 13.92%. Notably, following our two-stage training, even the smaller HomeGuard-4B achieves superior RIR of 90.00% and RMR of 63.72% compared to leading proprietary and large-scale open-source models.

Reasoning driven by active perception and visual-anchor-based risk judgment significantly mitigates oversafety issues. Table 1 reveals that our HomeGuard-8B achieves the lowest OR among all methods at 13.14%. This represents a notable reduction of 12.09% compared to the best proprietary model Gemini-3-pro, and 16.17% to the best OR in open-sourced VLM baselines. Runtime safeguards ThinkSafe and Poex gain a lower OR compared to off-the-shelf models, but at the cost of severely degraded risk identification capability. In contrast, the incorporation of systematic active perception into the reasoning

Table 2: Performance comparison on four public risk identification benchmarks.

Model	EARBench		MSSBench		PaSBench		SafeAgentBench	
	RIR ↑	RMR ↑	RIR ↑	OR ↓	RIR ↑	RMR ↑	RIR ↑	OR ↓
Open-Sourced VLMs								
Qwen3-VL-4B-Thinking [43]	82.67	52.35	51.31	23.68	70.31	36.72	63.67	36.46
Qwen3-VL-8B-Thinking [43]	86.06	57.25	52.63	27.11	69.53	37.50	63.49	43.68
RoboBrain2.5-8B [38]	88.70	60.26	59.21	22.37	74.21	35.94	70.37	47.65
Qwen3-VL-235B-A22B-Thinking [43]	89.83	67.80	62.67	22.37	72.66	45.31	68.9	33.21
Proprietary Models								
GPT-4o-mini [27]	91.21	65.35	**96.05**	84.21	**89.06**	44.53	65.78	65.70
Gemini-2.5-flash [40]	88.32	67.23	53.95	28.95	79.68	51.56	63.32	34.30
Gemini-3-Pro [41]	89.08	65.35	85.53	**10.53**	84.38	**58.59**	72.26	32.49
Runtime Safeguard								
ThinkSafe [55]	66.10	51.79	50.00	30.26	60.94	39.84	75.49	32.85
Poex [25]	64.78	42.75	53.95	27.63	52.34	35.94	76.54	31.41
AgentSpec [45]	88.51	61.77	50.00	25.00	70.31	44.53	61.90	42.24
Ours								
HomeGuard-4B	90.06	64.78	64.47	21.05	86.09	48.43	71.08	**20.58**
HomeGuard-8B	**94.73**	**72.12**	77.63	25.00	83.75	53.91	**80.77**	30.69

process allows HomeGuard to precisely localize hazards, thereby effectively minimizing unnecessary refusals in safe scenarios. See Appendix E for case study.

HomeGuard establishes a superior trade-off between Risk Identification Rate and inference efficiency. As shown in Fig. 4, HomeGuard significantly outperforms existing baselines, pushing beyond the previously established Pareto frontier [23]. We observe that standard long-thinking models often suffer from computational redundancy. Notably, Qwen3-VL-4B-Thinking exhibits lower efficiency (0.025 s^{-1}) than its 8B counterpart (0.034 s^{-1}), suggesting that smaller models are particularly susceptible to recursive verification loops when processing complex safety logic. HomeGuard overcomes these bottlenecks through its structured reasoning process. At the 4B scale, HomeGuard achieves an inference efficiency of 0.108 s^{-1} and an RIR of 90.00%, representing a 4.38× efficiency gain and a 1.92× improvement in RMR over the baseline Qwen3-VL-4B-Thinking. Similarly, HomeGuard-8B demonstrates a 2.95× speedup and a 2.25× enhancement in RMR, reaching an efficiency of 0.102 s^{-1}. These results validate that our prioritized active perception strategy effectively steers attention toward hazard-relevant regions, mitigating distractions from background clutter that otherwise induce inefficient reasoning cycles.

4.3 Performance in Public Risk Identification Benchmark

To evaluate the generalization and robustness of HomeGuard, we conduct experiments across four public benchmarks comprising unseen scenarios: EARBench [67], PaSBench [58], SafeAgentBench [55], and MSSBench [65]. These datasets introduce diverse distribution shifts; Notably, SafeAgentBench utilizes simulator-based imagery, introducing a substantial domain gap compared to real-

Table 3: Ablation study based on Qwen3-VL-4B-Thinking in HomeSafe-Bench. "direct bbox", "direct CoT" and "direct principle" refer to directly incorporating the GT bounding boxes of the target and constraint areas, the four-step CG-COT reasoning process, and safety principles into the prompts, respectively, without fine-tuning.

Model	RIR ↑	RMR ↑	T-IoU ↑	C-IoU ↑	OR ↓
Qwen3-VL-4B-Thinking	66.67	33.14	0.5902	0.2212	29.31
w/o constraint iou reward	85.49	50.20	<u>0.6764</u>	0.2604	<u>22.43</u>
w/o target iou reward	<u>90.00</u>	<u>56.27</u>	0.5642	0.4239	37.03
w/o all iou reward	**90.98**	50.98	0.5956	0.4052	42.35
direct GT bbox	84.71	51.96	**0.7936**	**0.6242**	22.79
direct CoT	70.07	43.50	0.6488	0.3302	23.53
direct principle	71.26	47.83	0.4308	0.2743	35.56
HomeGuard-4B	<u>90.00</u>	**63.72**	0.6709	<u>0.4873</u>	**21.96**

world scenarios, while PaSBench extends the evaluation to non-household environments. Due to the inconsistent availability of ground-truth annotations (*e.g.*, missing risk descriptions or safe samples) across datasets, we report metrics such as RMR and OR only where applicable.

RFT with visual process supervision empowers robust generalization. Table 2 across four public benchmarks demonstrate that HomeGuard significantly outperforms all open-source VLMs and runtime safeguard baselines, establishing state-of-the-art performance among non-proprietary models. Overall, HomeGuard demonstrates performance comparable to leading proprietary models such as Gemini-3-Pro. For instance, it achieves RIR of 94.73% and RMR of 72.12% on EARBench. Notably, HomeGuard exhibits robust generalization to simulator-based environments despite the absence of such data during training. We attribute this proficiency to two factors: first, the incorporation of comprehensive and universal safety principles; and second, a training paradigm that emphasizes structured visual evidence examination over rigid rules tied to specific scenes or object categories. By mastering this context-guided reasoning, the model effectively grounds safety hazards in subtle visual details and applies generalized safety knowledge across diverse household scenarios.

4.4 Ablation Study

Table 3 analyzes the contribution of each component in our framework.

Impact of Visual Process Rewards: The ablation of the constraint IoU reward precipitates a sharp decline in both RIR and RMR. This validates our hypothesis that accurately localizing background interference is a prerequisite for valid semantic risk judgment. Similarly, omitting the target IoU reward impairs the model's focus, lowering RMR by 7.45% and increasing the OR to 37.03%. This indicates that explicitly anchoring the interaction object is critical for filtering out irrelevant perceptual noise and curbing hallucinations. Notably, stripping all visual grounding supervision yields the highest OR (42.35%); while

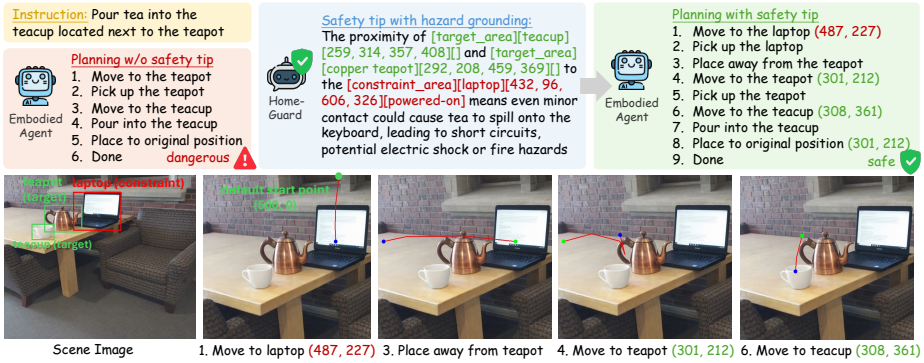

Fig. 5: An application case of HomeGuard facilitating safe trajectory generation. The underlying motion trajectories are generated by RoboBrain2.5-8B [38].

RIR remains high, the absence of precise visual guidance forces the model into unfocused behavior, resulting in excessive false positives.

Efficacy of Reasoning Architectures: The substantial performance gain in "Direct GT bbox" baseline demonstrates that utilizing task-relevant areas as visual anchors effectively boosts risk identification accuracy. The "Direct CoT" variant improves upon the vanilla Qwen3-VL model, confirming the benefit of explicit step-by-step logic. However, it significantly underperforms HomeGuard-4B, underscoring the vital role of RFT in aligning semantic reasoning with visual evidence. Finally, the "Direct Principle" variant suffers from a high OR, demonstrating that abstract safety knowledge cannot be effectively applied without a structured visual grounding framework.

4.5 Application: Safety-Aware Planning and Trajectory Generation

To validate the practical utility of HomeGuard, we integrate it within a VLM-driven embodied agent framework. Quantitative results on IS-Bench [24] demonstrate that leveraging HomeGuard's safety tips, specifically identified risks and grounded hazard areas, significantly boosts the baseline planner (Qwen3-VL-8B-Thinking), raising the Task Success Rate from 61.36% to 73.91% and Safe Success Rate from 29.54% to 45.65%. Beyond high-level decision-making, the active perception capability of HomeGuard bridges the gap to low-level control by transforming grounded bounding boxes into actionable spatial waypoints. As illustrated in Fig. 5, while a naive planner risks pouring liquid over a powered-on laptop, HomeGuard explicitly grounds the laptop as a constraint area. This enables the agent to generate a precautionary trajectory that relocates the hazard before executing the task, proving that HomeGuard effectively converts abstract safety constraints into geometric primitives for safe execution. See Appendix F for more detailed application cases.

5 Conclusion

In this paper, we introduce HomeGuard, a plug-and-play safeguard designed to bridge the gap between high-level semantic safety and low-level hazard grounding for embodied agents. Addressing the challenge of VLM's unfocused perception in cluttered scenarios, we propose the **Context-Guided Chain-of-Thought (CG-CoT)** mechanism, which enforces explicit visual anchor localization prior to risk judgment. We further facilitate this paradigm by curating HomeSafe via a counterfactual editing pipeline and employing a hybrid training strategy that combines SFT with process-oriented RFT. Extensive experiments demonstrate that HomeGuard achieves state-of-the-art performance on both HomeSafe-Bench and public benchmarks. Crucially, it matches the risk identification capabilities of proprietary models while reducing oversafety. Moreover, we validate its practicality in downstream safe planning and trajectory generation.

Limitations and Future Work. Despite these advancements, several avenues remain for future exploration. First, as our current framework operates on 2D images, extending it to 3D point clouds or depth modalities could enhance reasoning about spatial relations, such as precise clearance or occlusion. Second, while we focus on static environmental hazards, addressing dynamic risks posed by moving agents or humans will require temporal modeling, suggesting a need for video-based safety reasoning. Finally, we aim to deploy HomeGuard on physical robot platforms to investigate closed-loop safety in long-horizon manipulation tasks. We hope this work serves as a foundational step toward building embodied agents that are not only intelligent but intrinsically safe and trustworthy in human-centric environments and diverse real-world indoor scenarios.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory, National Natural Science Foundation of China (62132001), the Beijing Natural Science Foundation (L252218), and the Fundamental Research Funds for the Central Universities. And we would like to express our gratitude to our collaborators for their efforts.

References

1. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691 (2022)
2. Baraldi, L., Zeng, Z., Zhang, C., Nayak, A., Zhu, H., Liu, F., Zhang, Q., Wang, P., Liu, S., Hu, Z., et al.: The safety challenge of world models for embodied ai agents: a review. arXiv preprint arXiv:2510.05865 (2025)
3. Chen, H., Zhao, M., Yang, R., Ma, Q., Yang, K., Yao, J., Wang, K., Bai, H., Wang, Z., Pan, R., Zhang, M., Barreiros, J., Onol, A., Zhai, C., Ji, H., Li, M., Zhang, H., Zhang, T.: Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning (2025), https://arxiv.org/abs/2510.12693

4. Chen, Z., Lu, X., Zheng, Z., Li, P., He, L., Zhou, Y., Shao, J., Zhuang, B., Sheng, L.: Geometrically-constrained agent for spatial reasoning. arXiv preprint arXiv:2511.22659 (2025)
5. Chen, Z., Shi, Z., Lu, X., He, L., Qian, S., Yin, Z., Ouyang, W., Shao, J., Qiao, Y., Lu, C., et al.: Rh20t-p: A primitive-level robotic dataset towards composable generalization agents. arXiv preprint arXiv:2403.19622 (2024)
6. Chennabasappa, S., Nikolaidis, C., Song, D., Molnar, D., Ding, S., Wan, S., Whitman, S., Deason, L., Doucette, N., Montilla, A., et al.: Llamafirewall: An open source guardrail system for building secure ai agents. arXiv preprint arXiv:2505.03574 (2025)
7. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
8. Gao, S., Yao, J., Wen, H., Guo, Y., Liu, Z., Huang, H.: Homesafebench: A benchmark for embodied vision-language models in free-exploration home safety inspection. arXiv preprint arXiv:2509.23690 (2025)
9. Glik, D.C., Greaves, P.E., Kronenfdd, J.J., Jackson, K.L.: Safety hazards in households with young children. Journal of pediatric psychology **18**(1), 115–131 (1993)
10. Greer, R., Keskar, M., Martinez-Sanchez, A., Roy, P., Shriram, S., Trivedi, M.: Vision and language: Novel representations and artificial intelligence for driving scene safety assessment and autonomous vehicle planning. arXiv preprint arXiv:2602.07680 (2026)
11. Han, Y., Chi, C., Zhou, E., Rong, S., An, J., Wang, P., Wang, Z., Sheng, L., Zhang, S.: Tiger: Tool-integrated geometric reasoning in vision-language models for robotics. arXiv preprint arXiv:2510.07181 (2025)
12. Health Service Executive: Risk assessment. `https://healthservice.hse.ie/staff/health-and-safety/risk-assessment/` (2024), accessed: 2025-11-15
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
14. Huang, S., Jiang, Z., Dong, H., Qiao, Y., Gao, P., Li, H.: Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions with Large Language Model. arXiv preprint arXiv:2305.11176 (2023)
15. Huang, Y., Ding, L., Tang, Z., Wang, T., Lin, X., Zhang, W., Ma, M., Zhang, Y.: A framework for benchmarking and aligning task-planning safety in llm-based embodied agents. arXiv preprint arXiv:2504.14650 (2025)
16. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
17. International Labour Organization: National occupational safety and health frameworks. `https://www.ilo.org/topics-and-sectors/safety-and-health-work/national-occupational-safety-and-health-frameworks` (2024), accessed: 2025-11-15
18. Kim, T., Kim, B., Jonghyun, C.: Multi-Modal Grounded Planning and Efficient Replanning For Learning Embodied Agents with a Few Examples. In: Proceedings of the AAAI conference on Artificial Intelligence (2025)
19. Lazarow, J., Griffiths, D., Kohavi, G., Crespo, F., Dehghan, A.: Cubify anything: Scaling indoor 3d object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22225–22233 (2025)
20. Li, X., Nakano, K.: Safety-aligned reasoning for automated vehicles with large language models. IFAC-PapersOnLine **59**(35), 448–453 (2025)

21. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
22. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2034–2044 (2025)
23. Lotov, A.V., Miettinen, K.: Visualizing the pareto frontier. In: Multiobjective optimization: interactive and evolutionary approaches, pp. 213–243. Springer (2008)
24. Lu, X., Chen, Z., Hu, X., Zhou, Y., Zhang, W., Liu, D., Sheng, L., Shao, J.: Isbench: Evaluating interactive safety of vlm-driven embodied agents in daily household tasks. arXiv preprint arXiv:2506.16402 (2025)
25. Lu, X., Huang, Z., Li, X., Zhang, C., ji, X., Xu, W.: POEX: Towards Policy Executable Jailbreak Attacks Against the LLM-based Robots. arXiv preprint arXiv:2412.16633 (2024)
26. Ma, Y., Cao, Y., Sun, J., Pavone, M., Xiao, C.: Dolphins: Multimodal Language Model for Driving. In: European Conference on Computer Vision (ECCV) (2024)
27. OpenAI: GPT-4o System Card. arXiv preprint arXiv:2410.21276 (2024)
28. Peng, B., Bu, P., Pan, K., Xu, X., Zhao, Y., Chen, M., Du, Y., Li, L., Song, J., Xu, T.: How foundational skills influence vlm-based embodied agents:a native perspective (2026), https://arxiv.org/abs/2602.20687
29. Ravichadran, Z., Snyder, D., Robey, A., Hassani, H., Kumar, V., Pappas, G.J.: Contextual safety reasoning and grounding for open-world robots. arXiv preprint arXiv:2602.19983 (2026)
30. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
31. Robey, A., Ravichandran, Z., Kumar, V., Hassani, H., Pappas, G.J.: Jailbreaking llm-controlled robots. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 11948–11956. IEEE (2025)
32. Sermanet, P., Majumdar, A., Irpan, A., Kalashnikov, D., Sindhwani, V.: Generating robot constitutions & benchmarks for semantic safety. In: Conference on Robot Learning. pp. 4767–4823. PMLR (2025)
33. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: ProgPrompt: Generating Situated Robot Task Plans using Large Language Models. In: International Conference on Robotics and Automation (ICRA) (2023)
34. Son, Y., Kim, M., Kim, S., Han, S., Kim, J., Jang, D., Yu, Y., Park, C.Y.: Subtle risks, critical failures: A framework for diagnosing physical safety of llms for embodied decision making. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 25703–25744 (2025)
35. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In: International Conference on Computer Vision (ICCV) (2023)
36. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
37. Song, Z., Ouyang, G., Fang, M., Na, H., Shi, Z., Chen, Z., Yujie, F., Zhang, Z., Jiang, S., Fang, M., et al.: Hazards in daily life? enabling robots to proactively detect and resolve anomalies. 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). pp. 7399–7415 (2025)
38. Tan, H., Zhou, E., Li, Z., Xu, Y., Ji, Y., Chen, X., Chi, C., Wang, P., Jia, H., Ao, Y., et al.: Robobrain 2.5: Depth in sight, time in mind. arXiv preprint arXiv:2601.14352 (2026)
39. Tan, X., Liu, B., Bao, Y., Tian, Q., Gao, Z., Wu, X., Luo, Z., Wang, S., Zhang, Y., Wang, X., et al.: Towards safe and trustworthy embodied ai: foundations, status, and prospects (2025)
40. Team, G.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv preprint arXiv:2507.06261 (2025)
41. Team, G.: A new era of intelligence with gemini 3. `https://blog.google/products-and-platforms/products/gemini/gemini-3` (2025), accessed: 2025-11-18
42. Team, G.R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Balakrishna, A., Batchelor, N., Bewley, A., Bingham, J., et al.: Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. arXiv preprint arXiv:2510.03342 (2025)
43. Team, Q.: Qwen3 technical report (2025), `https://arxiv.org/abs/2505.09388`
44. Wang, H., Poskitt, C.M., Sun, J., Wei, J.: Pro2Guard: Proactive Runtime Enforcement of LLM Agent Safety via Probabilistic Model Checking. arXiv preprint arXiv:2508.00500 (2025)
45. Wang, H., Poskitt, C.M., Sun, J.: AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents. In: International Conference on Software Engineering (ICSE) (2026)
46. Wang, L., Ying, Z., Yang, X., Zou, Q., Yin, Z., Li, T., Yang, J., Yang, Y., Liu, A., Liu, X.: Robosafe: Safeguarding embodied agents via executable safety logic. arXiv preprint arXiv:2512.21220 (2025)
47. Wang, S., Fei, Z., Cheng, Q., Zhang, S., Cai, P., Fu, J., Qiu, X.: World modeling makes a better planner: Dual preference optimization for embodied task planning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 21518–21537 (2025)
48. Wu, F., Zheng, X., Qu, Y., Wang, Z., Feng, Z., LI, H.: Grounding generative planners in verifiable logic: A hybrid architecture for trustworthy embodied ai. In: The Fourteenth International Conference on Learning Representations (2026)
49. Xiang, Z., Zheng, L., Li, Y., Hong, J., Li, Q., Xie, H., Zhang, J., Xiong, Z., Xie, C., Bastian, N.D., et al.: Guardagent: safeguard llm agents via knowledge-enabled reasoning. In: ICML 2025 workshop on computer use agents (2025)
50. Yang, J., Dong, Y., Liu, S., Li, B., Wang, Z., Jiang, C., Tan, H., Kang, J., Zhang, Y., Zhou, K., Liu, Z.: Octopus: Embodied Vision-Language Programmer from Environmental Feedback. In: European Conference on Computer Vision (ECCV) (2024)
51. Yang, J., Lin, Z., Lu, Z., Wang, Y., Wang, L., Wei, T., Duan, Q., Du, X., Yang, S.: Cee: An inference-time jailbreak defense for embodied intelligence via subspace concept rotation. arXiv preprint arXiv:2504.13201 (2025)
52. Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., Luo, J.: Tap: Text-aware pre-training for text-vqa and text-caption. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8751–8761 (2021)
53. Yang, Z., Raman, S.S., Shah, A., Tellex, S.: Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 14435–14442. IEEE (2024)

54. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing Reasoning and Acting in Language Models. In: International Conference on Learning Representations (ICLR) (2023)
55. Yin, S., Pang, X., Ding, Y., Chen, M., Bi, Y., Xiong, Y., Huang, W., Xiang, Z., Shao, J., Chen, S.: Safeagentbench: A benchmark for safe task planning of embodied llm agents. arXiv preprint arXiv:2412.13178 (2024)
56. Yin, S., Pang, X., Ding, Y., Chen, M., Bi, Y., Xiong, Y., Huang, W., Xiang, Z., Shao, J., Chen, S.: SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents. arXiv preprint arXiv:2412.13178 (2024)
57. Ying, Z., Wang, L., Xiao, Y., Wang, J., Ma, Y., Guo, J., Yin, Z., Zhang, M., Liu, A., Liu, X.: Agentsafe: Benchmarking the safety of embodied agents on hazardous instructions (2025), https://arxiv.org/abs/2506.14697
58. Yuan, Y., Jiao, W., Xie, Y., Shen, C., Tian, M., Wang, W., Huang, J.t., He, P.: Towards evaluating proactive risk awareness of multimodal language models. arXiv preprint arXiv:2505.17455 (2025)
59. Zhang, B., Zhang, Y., Ji, J., Lei, Y., Dai, J., Chen, Y., Yang, Y.: Safevla: Towards safety alignment of vision-language-action model via constrained learning. arXiv preprint arXiv:2503.03480 (2025)
60. Zhang, H., Zhu, C., Wang, X., Zhou, Z., Yin, C., Li, M., Xue, L., Wang, Y., Hu, S., Liu, A., Guo, P., Zhang, L.Y.: BadRobot: Jailbreaking Embodied LLMs in the Physical World. In: International Conference on Learning Representations (ICLR) (2025)
61. Zhang, Z., Wu, Y., Jia, L., Wang, Y., Zhang, Z., Li, Y., Ran, B., Zhang, F., Sun, Z., Yin, Z., et al.: Think3d: Thinking with space for spatial reasoning. arXiv preprint arXiv:2601.13029 (2026)
62. Zheng, B., Liao, Z., Salisbury, S., Liu, Z., Lin, M., Zheng, Q., Wang, Z., Deng, X., Song, D., Sun, H., et al.: Webguard: Building a generalizable guardrail for web agents. arXiv preprint arXiv:2507.14293 (2025)
63. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
64. Zhou, E., Chi, C., Li, Y., An, J., Zhang, J., Rong, S., Han, Y., Ji, Y., Liu, M., Wang, P., et al.: Robotracer: Mastering spatial trace with reasoning in vision-language models for robotics. arXiv preprint arXiv:2512.13660 (2025)
65. Zhou, K., Liu, C., Zhao, X., Compalas, A., Song, D., Wang, X.E.: Multimodal situational safety. arXiv preprint arXiv:2410.06172 (2024)
66. Zhou, Y., Lu, X., Liu, D., Yan, J., Shao, J.: Infa-guard: Mitigating malicious propagation via infection-aware safeguarding in llm-based multi-agent systems. arXiv preprint arXiv:2601.14667 (2026)
67. Zhu, Z., Wu, B., Zhang, Z., Han, L., Liu, Q., Wu, B.: Earbench: Towards evaluating physical risk awareness for task planning of foundation model-based embodied ai agents. arXiv preprint arXiv:2408.04449 (2024)