基于 **Gemini Robotics 1.5** 的通用机器人模型在人工智能前沿中的应用

曾康慧

2026 年 1 月 9 日

摘　要

通用机器人是人工智能前沿领域的重要发展方向，需要对物理世界具备深度理解、高级推理能力和灵活的动作控制。本文基于 Google DeepMind 发布的 Gemini Robotics 1.5 模型家族，系统探讨其在多模态感知、具身推理（Embodied Reasoning, ER）和视觉-语言-动作（Vision-Language-Action, VLA）控制方面的核心创新。该模型通过引入运动转移（Motion Transfer, MT）机制、具身思考范式以及先进的代理架构，实现了跨机器人平台的零样本技能迁移、复杂长时序任务的分解执行以及更高的安全性和可解释性。本文首先回顾相关工作，随后详细介绍模型架构、训练方法与关键技术，继而通过大量实验结果分析其在短时序泛化、多实施体学习、具身推理以及代理系统中的性能表现，并讨论安全与责任开发实践。最后，本文对通用机器人未来发展进行展望，为人工智能前沿课程提供一个全面、深入的案例研究。

关键词：通用机器人；具身推理；视觉语言动作模型；运动转移；多实施体学习；人工智能前沿

目录

1 引言 **1**

2 相关工作 **2**

2.1 视觉语言模型在机器人控制中的应用 2

2.2 多实施体学习与技能迁移 2

2.3 具身推理与代理系统 2

2.4 安全与责任开发 3

3 模型架构与方法概述 **3**

3.1 整体框架 3

3.2 运动转移机制 3

3.3 具身思考范式 3

3.4 训练数据与评估方法 3

4 **Gemini Robotics 1.5 VLA** 模型性能分析 **4**

4.1 多机器人任务示例 4

4.2 短时序泛化能力 4

4.3 多实施体学习消融 4

4.4 具身思考对多步任务的影响 6

5 **Gemini Robotics-ER 1.5** 具身推理能力 **6**

6 代理系统性能 **7**

7 安全与责任开发 **7**

8 结论与展望 **8**

1 引言

随着人工智能技术的迅猛发展，机器人正从专用任务执行者向通用智能体演进。真正的通用机器人需要对物理世界具备深刻的理解，能够进行高级推理，并实现灵活、灵巧的动作控制。然而，传统机器人学习面临三大瓶颈：数据稀缺与异构性、泛化能力不足以及长时序复杂任务的规划与执行困难。

Google DeepMind 于 2025 年发布的 Gemini Robotics 1.5 模型家族为上述问题提供了系统性解决方案。该模型家族包括 Gemini Robotics 1.5（视觉语言动作模型，VLA）和 Gemini Robotics-ER 1.5（具身推理模型），并通过代理架构将二者有机结合。如图1所示，该框架实现了感知、思考与行动的闭环，显著提升了机器人在真实世界中的鲁棒性和智能水平。

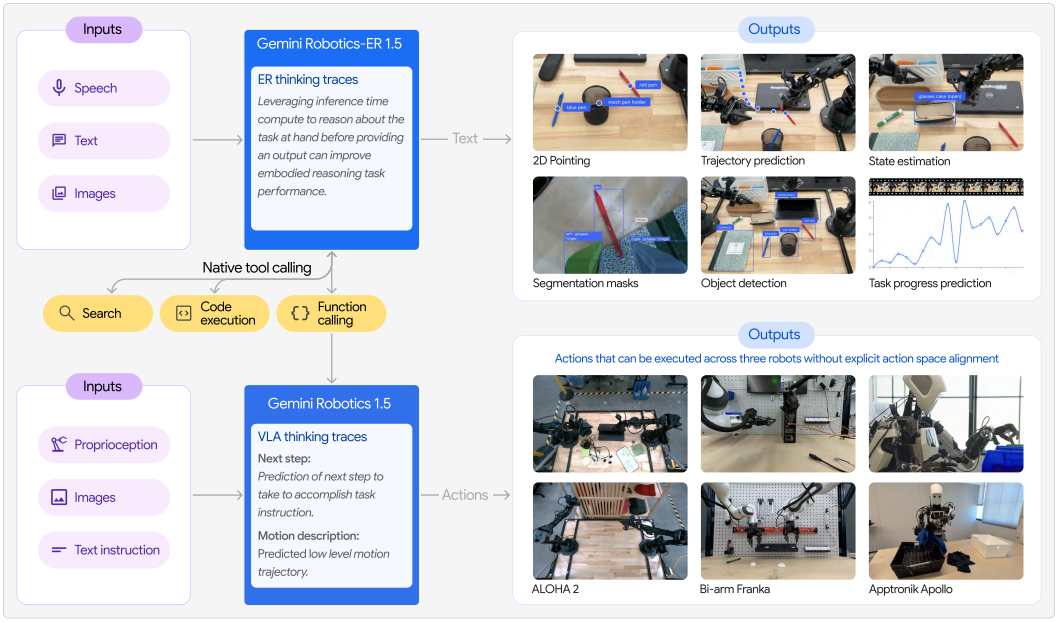


图 1 Gemini Robotics 1.5 整体框架。如图所示，该系统由具身推理模型（ER）作为高层协调器、视觉语言动作模型（VLA）作为低层执行器构成，支持多模态交互、工具调用以及多级具身思考。

本文基于 Gemini Robotics 1.5 技术报告，对其核心创新进行系统分析。首先回顾相关工作，随后详细阐述模型架构与方法，继而通过实验结果分析其在泛化、推理与代理性能方面的表现，最后展望未来发展方向。本文旨在为“人工智能前沿”课程提供一个典型案例，帮助理解多模态大模型如何推动具身智能的进步。

2 相关工作

2.1 视觉语言模型在机器人控制中的应用

视觉语言模型（Vision-Language Models, VLMs）为机器人提供了强大的语义理解能力。RT-2 模型[1]首次将网络规模的视觉语言知识迁移至机器人控制，展示了 VLMs 在开放词汇指令理解上的优势。随后，Gemini 系列模型[2]进一步提升了多模态推理、长上下文处理以及代理能力，为 Gemini Robotics 1.5 奠定了基础。

在具身任务中，VLMs 常被用作成功检测器[3]或零样本奖励模型[4]，为强化学习提供语义指导。这些工作表明，语言空间的推理能够显著简化机器人策略学习。

2.2 多实施体学习与技能迁移

机器人数据通常高度异构，不同平台（如双臂、移动机械臂、人形机器人）在动作空间与感知方式上差异巨大。早期工作依赖模拟器（如 MuJoCo[5]）进行大规模评估，而近期 GR00T[6]、DexVLA[7]等模型尝试从多平台数据中学习通用策略。

Gemini Robotics 1.5 引入的运动转移（Motion Transfer, MT）机制是该方向的重要突破。通过新型架构与训练配方，模型能够在不进行平台特定后训练的情况下，直接控制 ALOHA、Bi-arm Franka 以及 Apollo 人形机器人，并实现零样本技能跨平台迁移。

2.3 具身推理与代理系统

具身推理强调对物理世界的空间、时序与因果理解。Robo Spatial[8]和 Point Arena[9]等基准推动了空间推理能力的发展。交互式任务规划[10]与错误恢复[11]研究则探索了语言模型在高层规划中的作用。

代理系统方面，SayCan[12]与 Inner Monologue[13]展示了将大语言模型与低层策略结合的可行性。Gemini Robotics 1.5 在此基础上引入具身思考范式，使模型能够在动作前生成自然语言内部独白，从而提升任务分解与可解释性。

2.4 安全与责任开发

机器人安全涉及语义约束、物理碰撞避免以及社会影响。ASIMOV 基准系列[14]为语义动作安全提供了评估框架。ISO/TS 15066:2016 与 ISO 10218-1:2025 等标准规范了协作机器人安全要求。Gemini Robotics 1.5 通过多层安全机制（包括自动红队）显著提升了部署可靠性。

3 模型架构与方法概述

3.1 整体框架

Gemini Robotics 1.5 模型家族继承了 Gemini 的多模态世界知识。如图1所示，系统包含具身推理模型（ER）与视觉语言动作模型（VLA）两大核心组件，通过代理架构实现协同。

3.2 运动转移机制

运动转移（MT）是实现多实施体学习的关键创新。模型通过统一运动表示，将源机器人状态-动作对映射至目标机器人：

$$f_\theta : (S_r, A_r) \mapsto (S_t, A_t)$$

其中 θ 为共享参数。该机制使得模型能够在异构数据上预训练，实现跨平台零样本控制。

3.3 具身思考范式

具身思考允许模型在动作前生成多级自然语言思考轨迹。这种“思考后行动”的方式将复杂指令分解为短时序子目标，提升了长时序任务的成功率与可解释性。

3.4 训练数据与评估方法

训练数据涵盖多平台真实轨迹以及互联网文本、图像、视频。评估采用真实机器人 A/B 测试，并通过 MuJoCo 模拟器加速迭代。如图2所示，模拟与真实评估具有高度排名一致性。

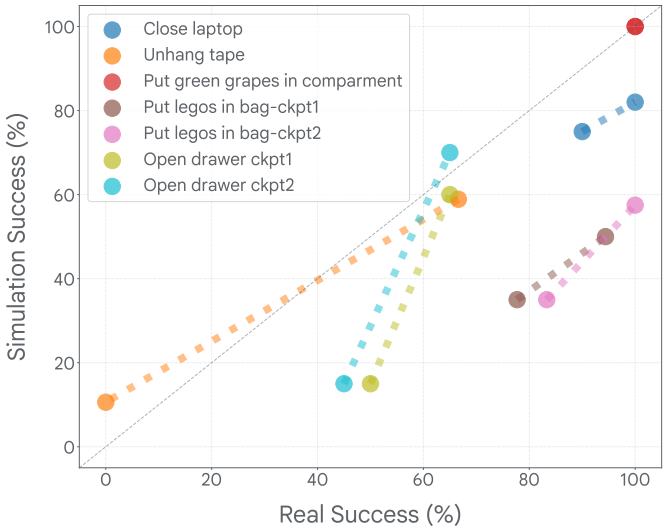


图 2 模拟到真实的相关性。如图所示，成功率在模拟与真实环境中的排名高度一致，支持快速迭代。

4 Gemini Robotics 1.5 VLA 模型性能分析

4.1 多机器人任务示例

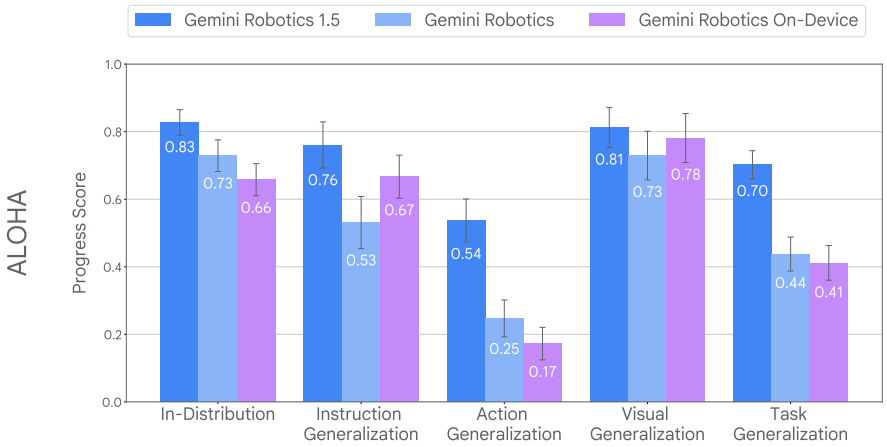
图 3 多机器人任务示例。如图所示，同一模型无需后训练即可控制 ALOHA、Bi-arm Franka 以及 Apollo 人形机器人完成多样化任务。

4.2 短时序泛化能力

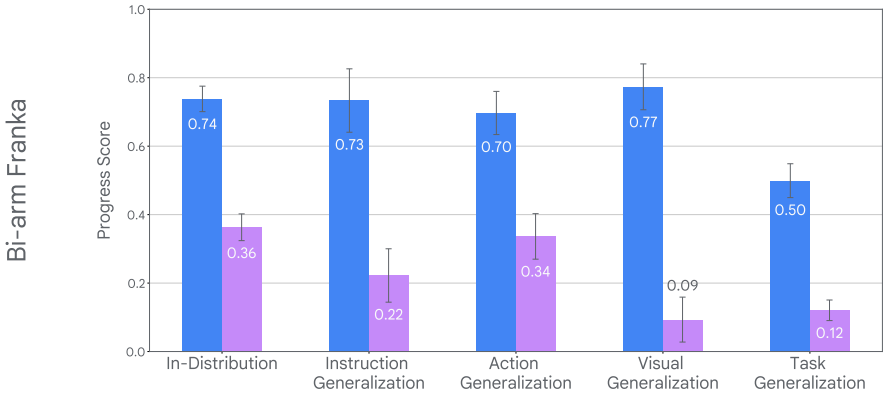
模型在视觉、语义、动作和新任务四个维度上均表现出色。如图4a、4b所示，Gemini Robotics 1.5 显著优于基线。

4.3 多实施体学习消融

如图5所示，多平台数据结合 MT 配方带来显著正向迁移。



(a) ALOHA 平台



(b) Bi-arm Franka 平台

图 4 短时序泛化性能分解

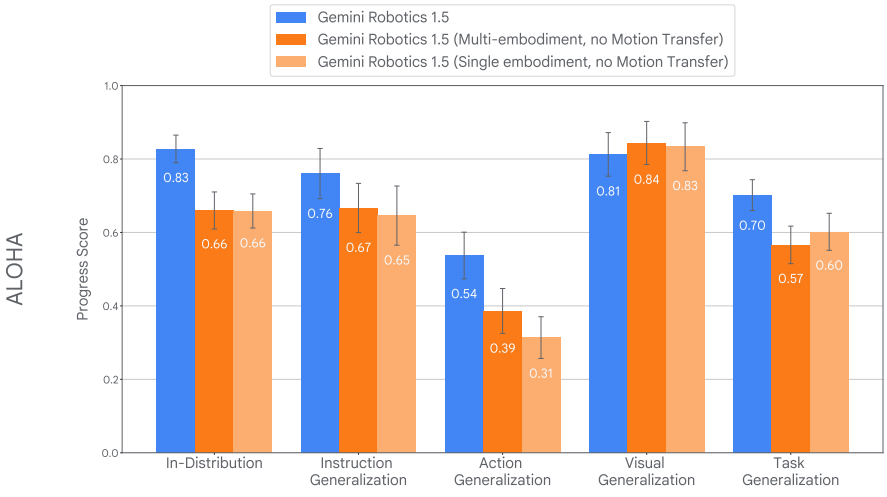


图 5 多实施体学习消融实验

4.4 具身思考对多步任务的影响

启用思考模式后，多步任务性能大幅提升。如图6所示，进步分数提升约 20%。

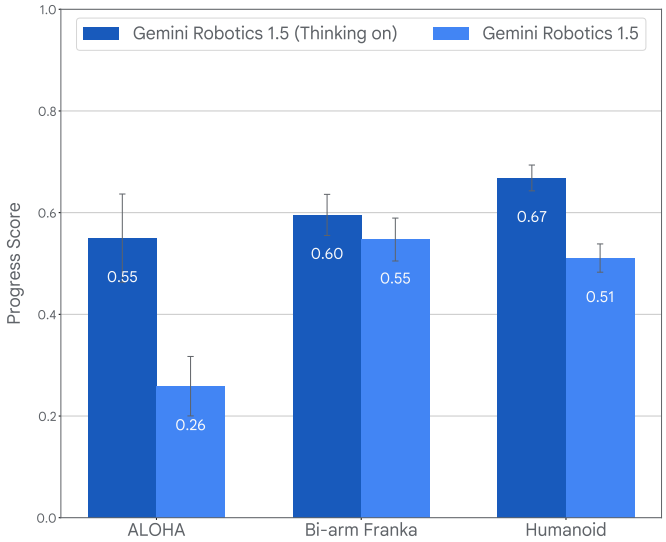


图 6 具身思考消融

5 Gemini Robotics-ER 1.5 具身推理能力

如图7所示，该模型在保持前沿模型通用性的同时，在具身推理基准上取得 state-of-the-art。

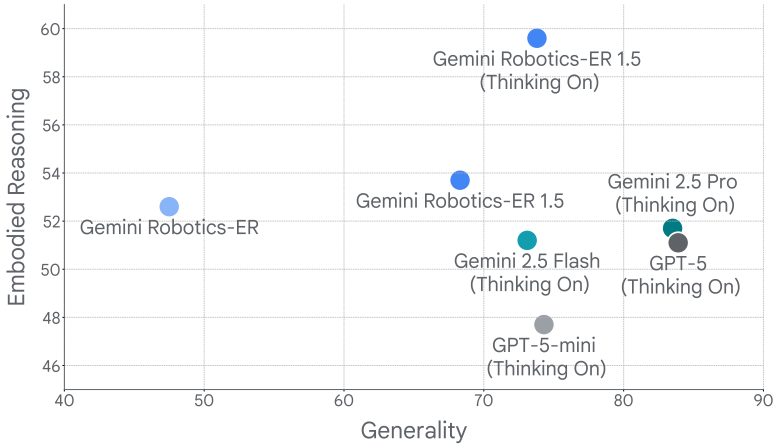


图 7 通用性与具身推理权衡

如图8所示，模型在指向、进度估计、视频问答等多项机器人关键能力上表现突出。

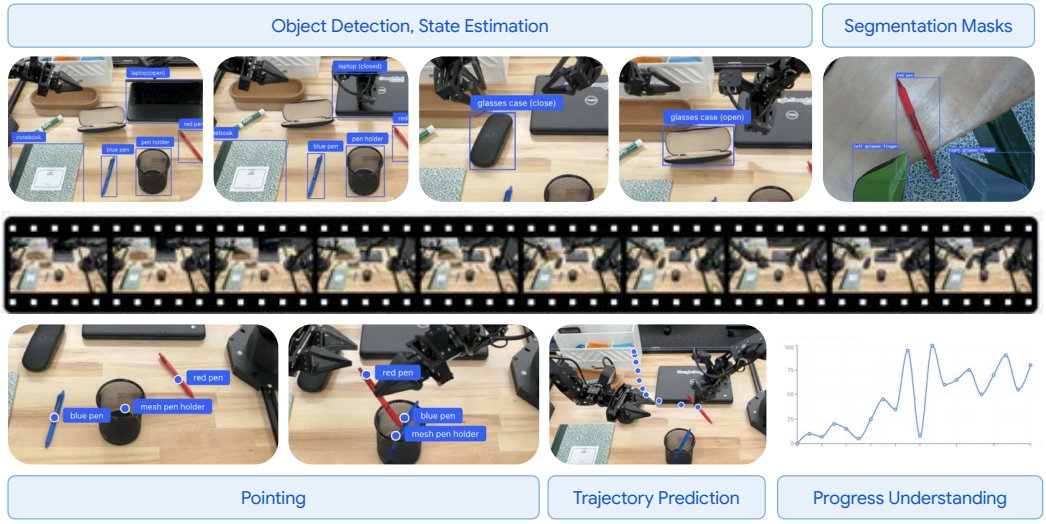


图 8 具身推理关键能力分解

6 代理系统性能

将 ER 与 VLA 结合构成完整代理后，长时序任务成功率显著提升。如图9所示，Gemini Robotics 1.5 代理在复杂任务中接近 80% 进步分数。

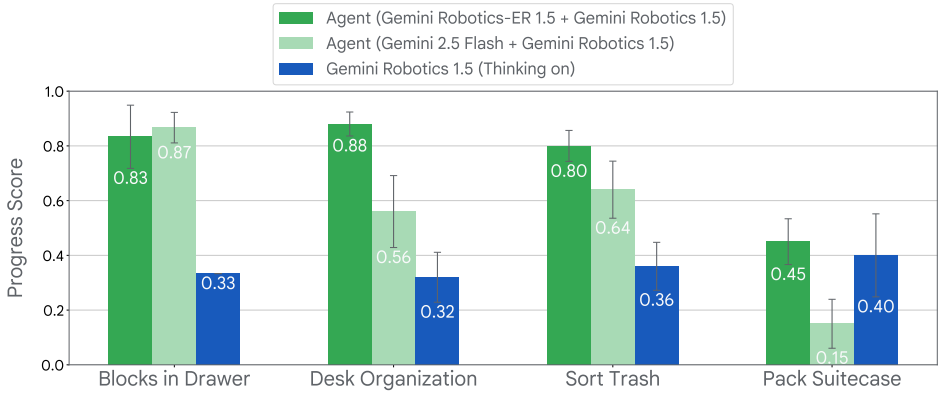
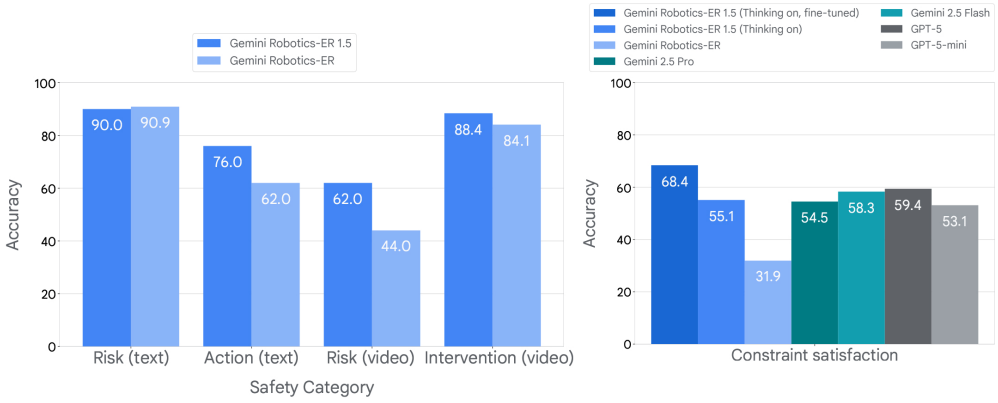


图 9 长时序代理性能对比

失败模式分析（如表**??**）表明，ER 模型大幅降低了规划错误率。

7 安全与责任开发

模型采用多层安全机制，包括语义动作安全与自动红队框架。如图10所示，在 ASIMOV-2.0 基准上取得 state-of-the-art。



(a) GR-ER 1.5 improves injury risk recognition, action safety and intervention prediction.

(b) GR-ER 1.5 Thinking improves adherence to Physical Safety Constraints.

图 10 ASIMOV-2.0 安全评估结果

图 11 自动红队检测 ER 幻觉示例

8 结论与展望

Gemini Robotics 1.5 标志着通用机器人研究的重要里程碑。其三大核心创新——运动转移机制、具身思考范式以及先进的具身推理能力——共同解决了数据异构、泛化不足与长时序规划三大难题。实验结果表明，该模型不仅在短时序任务中展现出卓越的跨平台泛化能力，还在长时序复杂场景中通过代理架构实现了接近人类水平的智能表现。同时，多层安全机制为其实际部署提供了可靠保障。

展望未来，通用机器人仍有广阔发展空间。首先，可进一步整合更多模态数据（如真实人类视频、合成视频以及音频、触觉信号），缓解数据稀缺问题。其次，将强化学习与 VLA 预训练结合，有望在保持泛化性的同时显著提升灵巧性。第三，探索轮式、飞行等新型实施体，扩展 MT 机制的应用范围。第四，加强伦理与社会影响研究，制定公平部署政策，缓解就业冲击与隐私风险。第五，推动部分组件开源，促进全球研究社区协作。

总之，Gemini Robotics 1.5 不仅为当前机器人学习提供了实用范式，也为未来具身智能的全面实现指明了方向。随着计算能力与数据规模的持续增长，我们有理由相信，真正融入物理世界的通用人工智能代理即将到来。

参考文献

[1] ZITKOVICH B, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[M]//CoRL. 2023.

[2] COMANICI G, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities[A]. 2025.

[3] DU Y, et al. Vision-language models as success detectors[A]. 2023.

[4] ROCAMONDE J, et al. Vision-language models are zero-shot reward models for reinforcement learning[A]. 2023.

[5] TODOROV E, EREZ T, TASSA Y. Mujoco: A physics engine for model-based control[C]//IROS. 2012.

[6] BJORCK J, et al. Gr00t n1: An open foundation model for generalist humanoid robots[A]. 2025.

[7] WEN J, et al. Dexvla: Vision-language model with plug-in diffusion expert for general robot control[A]. 2025.

[8] SONG C H, et al. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics[C]//CVPR. 2025.

[9] CHENG L, et al. Pointarena: Probing multimodal grounding through language-guided pointing[A]. 2025.

[10] LI B, et al. Interactive task planning with language models[A]. 2023.

[11] CHEN H, et al. Automating robot failure recovery using vision-language models with optimized prompts[A]. 2024.

[12] AHN M, et al. Do as i can, not as i say: Grounding language in robotic affordances [J]. CoRL, 2022.

[13] HUANG W, et al. Inner monologue: Embodied reasoning through planning with language models[A]. 2022.

[14] SERMANET P, et al. Generating safety benchmarks for vision-language models in robotics[A]. 2025.