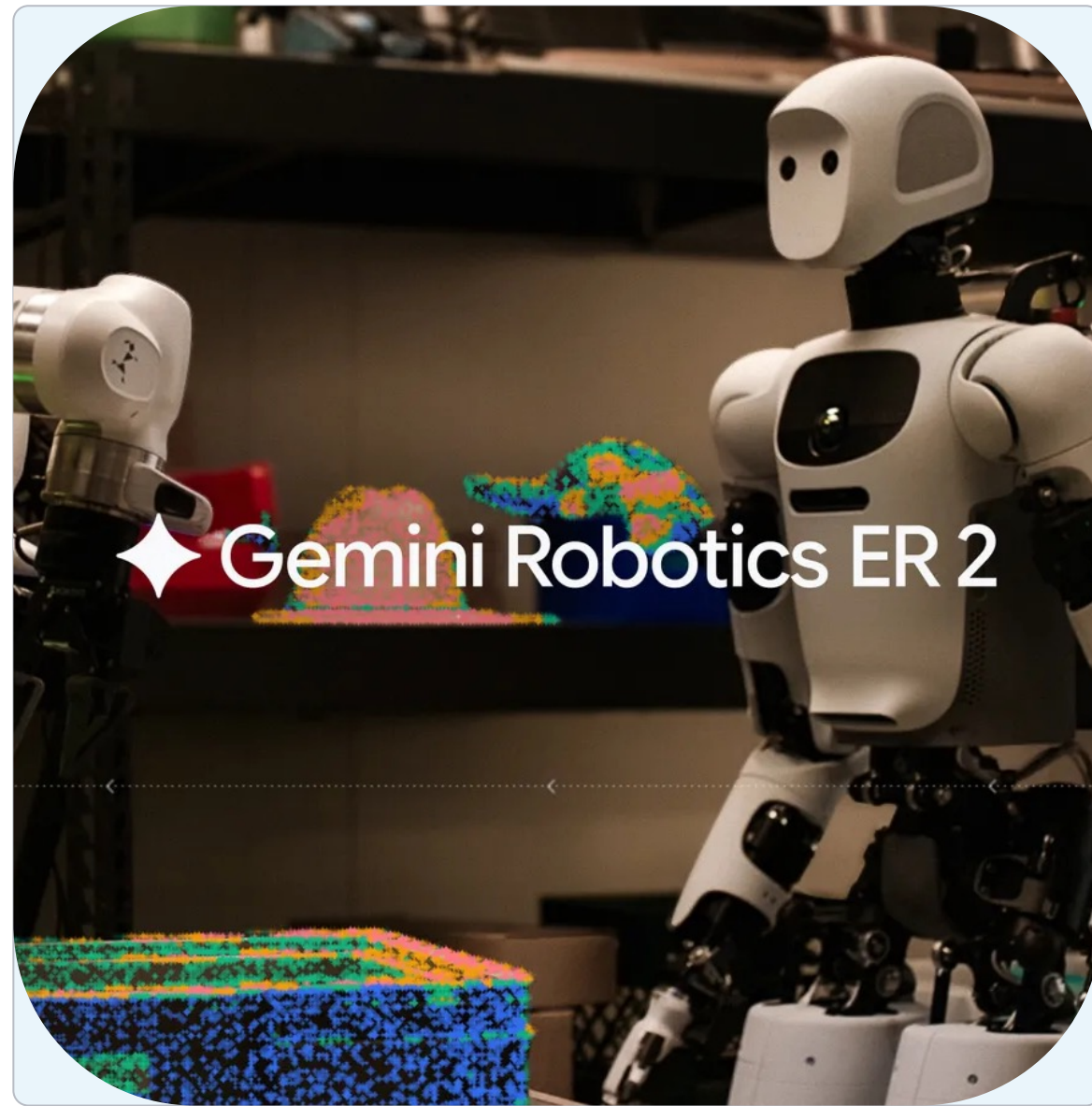


谷歌机器人的 “高层大脑”

架构转折、平台机会与落地边界

洞察材料 | 2026.08



ER 2 不是 VLA 升级，而是机器人系统的上层补位

两类模型解决不同时间尺度的问题，正确关系是协同，不是替代。

ER 2 / 认知控制层

做什么、做到哪一步、是否成功

连续视频理解
多步骤规划与失败重规划
VLA / 导航 / 机器人 API 调度

VLA / 动作执行层

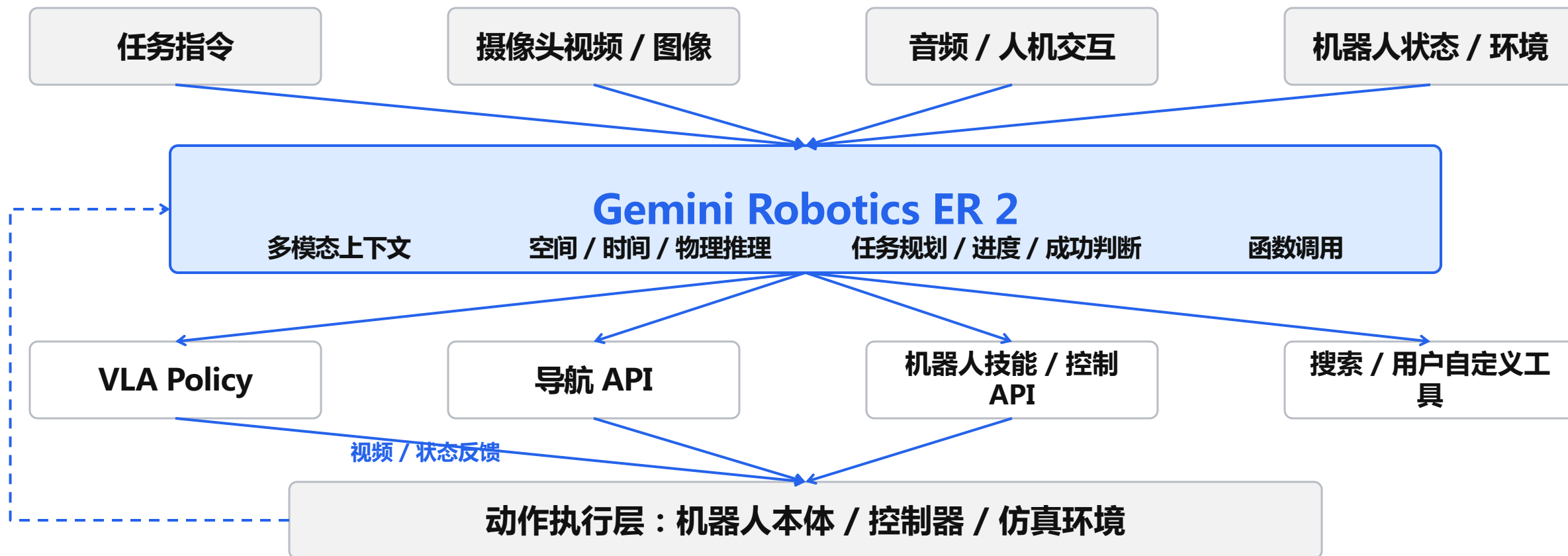
下一段动作如何执行

输出关节动作或末端位姿
处理毫秒到数秒的局部控制
完成抓取、放置、导航等技能

系统结论：ER 2 决定“何时调用哪种能力”，VLA 决定“动作怎么做”。

ER 2 技术方案：四层架构解耦推理与控制

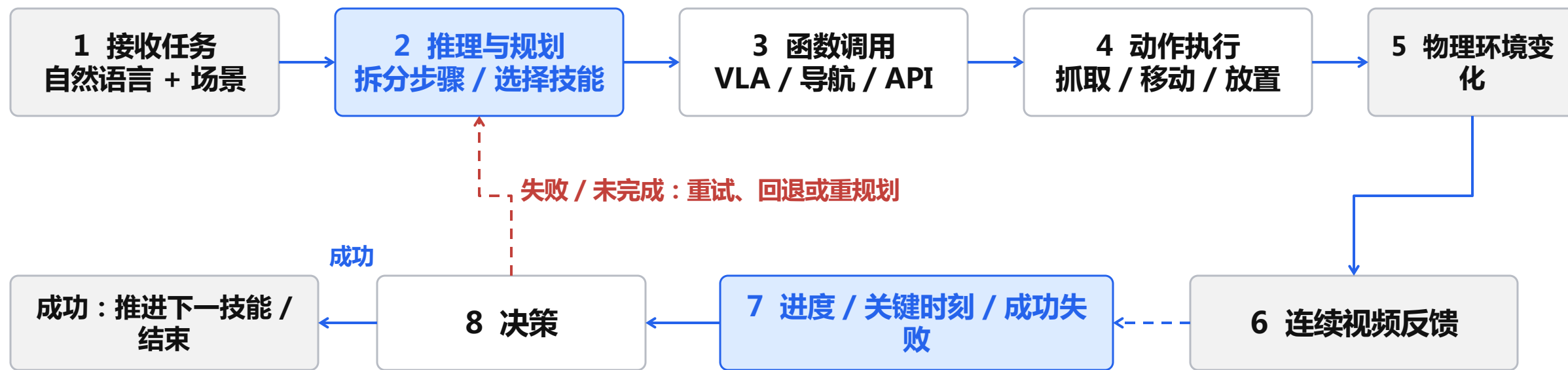
Gemini 3.5 Flash | 131K 输入上下文 | 文本 / 图像 / 视频 / 音频 → 文本与函数调用



独立安全底座：急停、碰撞检测、力矩限制、安全区域与安全 PLC 对动作保留最终否决权

闭环执行：每次动作都经过观察、判断与校正

ER 2 运行在低频认知闭环，底层 VLA 与控制器运行在高频动作闭环。



关键接口：ER 2 输出计划、坐标与函数调用；VLA / 控制器输出关节动作、末端位姿或 action chunk。

工程接入：标准端点与 Streaming 端点各司其职

两个端点共享多模态输入与函数调用能力，但面向不同延迟与工具需求。

标准端点

`gemini-robotics-er-2-preview`

能力

代码执行、结构化输出、文件搜索
Function Calling、搜索与空间推理

适用

复杂任务规划、空间分析、离线评测

Streaming 端点

`gemini-robotics-er-2-streaming-preview`

能力

Live API 双向音视频流、低延迟函数调用
不支持代码执行、结构化输出与文件搜索

适用

实时任务编排、人机交互、连续视频监督

推荐组合：Streaming 端点承担在线闭环，标准端点承担复杂规划、结构化分析与离线复盘。

真正的升级：三项能力第一次形成闭环

长时程规划 × 连续视频状态估计 × 工具与机器人调度



01 规划

把自然语言任务拆成多步骤，并在失败后重新组织计划。



02 监督

持续判断进度、关键时刻与成功失败，让执行真正闭环。



03 调度

把 VLA、导航和不同机器人抽象为工具，按能力动态调用。

洞察：价值不是单项 VQA 变强，而是系统模块之间的断点减少。

时间判断成为机器人系统的新控制变量

机器人常见问题不是“不知道怎么做”，而是不知道何时继续、何时停止、何时重试。

57.4%

进度分类准确率
逐帧识别五档任务进度

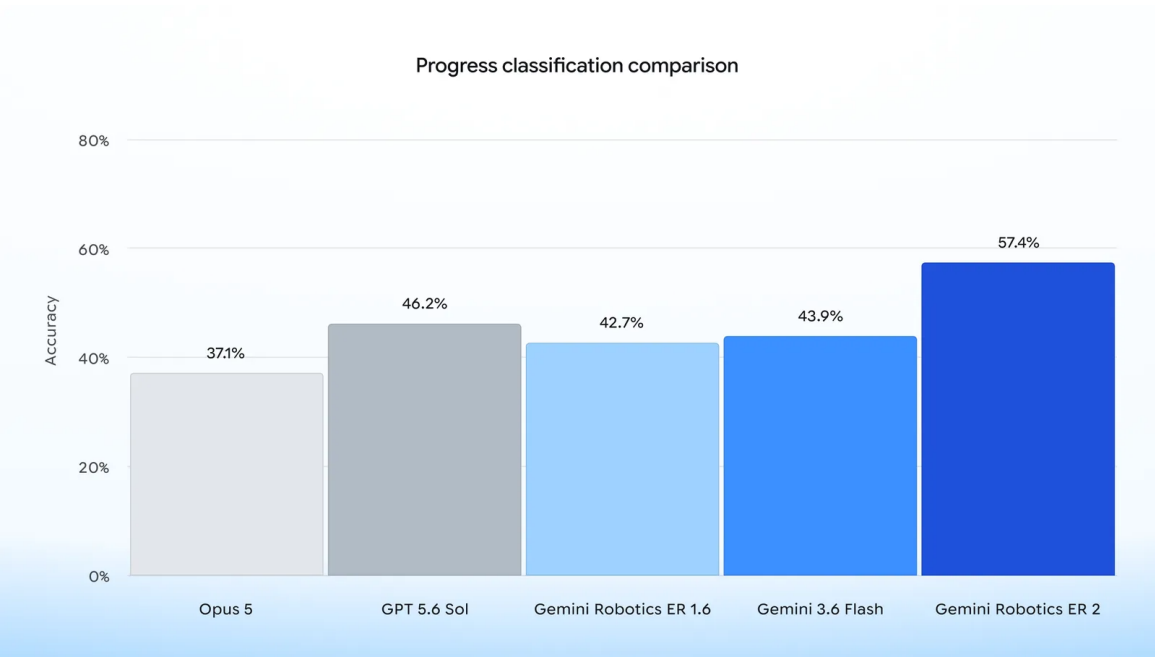
91.3%

关键时刻定位准确率
识别事件发生的精确帧

0.96 秒

关键时刻平均绝对误差
让技能切换更接近实时

指标已证明价值，但尚未证明可独立托底



57.4% 是相对领先，
不是生产就绪证书

领先

相对 ER 1.6 与对比模型
进度理解显著提升

Preview

仍存在幻觉、延迟与
分布外输入风险

多机器人协作，本质是共享语义而非共享动作

不同 embodiment 不必统一关节结构，只需把能力、状态和任务交接放到同一语义层。

能力抽象

将 Apollo 的移动取物、Franka Duo 的精细操作抽象为技能。

任务交接

ER 2 按场景把任务从一个机器人移交给另一个机器人。

统一验收

通过多路视频反馈判断整体工作流是否真正完成。

产业含义：高层 Agent 可以跨本体复用，底层 VLA 继续保留各自最优动作表示。

对 VLA/WAM 平台，最短路径是监督与评测

先利用 ER 2 的强项建立可量化闭环，再考虑让它承担更大范围的调度权。

01 VLA 执行监督

连续视频 → 进度 / 成功 / 失败

用于抓取、插入、摆放等任务的自动验收；失败时给出重试或回退信号。

02 WAM 评测器

预测视频 ↔ 真实视频

比较关键事件、任务进度和失败类型，让世界模型评测从像素差异走向任务语义。

第三阶段再扩展到多机器人调度：先验证判断可靠性，再放大系统权限。

机会明确，但 Preview 边界必须前置

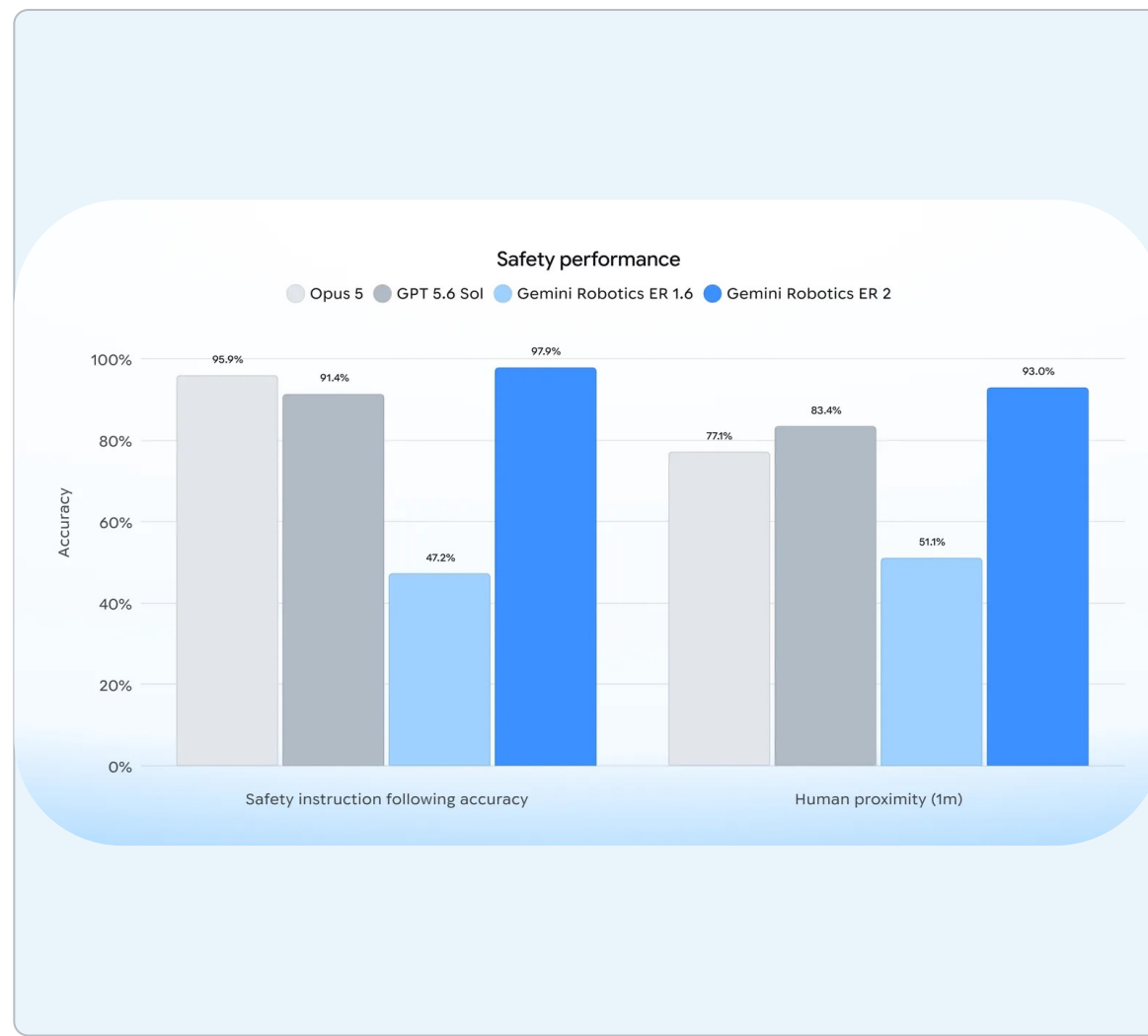
适合先做

执行监督与自动验收
仿真 / 真机视频评测
非关键流程的工具调度

不能替代

高频运动控制与力觉闭环
独立急停、碰撞检测、力矩限制
安全区域、人体检测与安全 PLC

部署原则：模型负责认知建议，传统安全系统保留最终否决权。



建议

先证明 “看得准” 再扩大 “管得多”

P0 VLA 执行监督：进度、成功、失败是否稳定

P1 WAM 语义评测：预测视频是否保留关键事件与任务因果

P2 多机器人调度：在保留安全底座的前提下扩展权限