Position: Modular Safety Guardrails Are Necessary for Foundation-Model-Enabled Robots in the Real World

Joonkyung Kim [1 †] **Wenxi Chen** [2 †] **Davood Soleymanzadeh** [1 †] **Yi Ding** [2 ‡] **Xiangbo Gao** [1 ‡] **Zhengzhong Tu** [1]
Ruqi Zhang [2] **Fan Fei** [3] **Sushant Veer** [4] **Yiwei Lyu** [1 *] **Minghui Zheng** [1 *] **Yan Gu** [2 *]

Abstract

The integration of foundation models (FMs) into robotics has accelerated real-world deployment, while introducing new safety challenges arising from open-ended semantic reasoning and embodied physical action. These challenges require safety notions beyond physical constraint satisfaction. In this position paper, we characterize FM-enabled robot safety along three dimensions: action safety (physical feasibility and constraint compliance), decision safety (semantic and contextual appropriateness), and human-centered safety (conformance to human intent, norms, and expectations). We argue that existing approaches, including static verification, monolithic controllers, and end-to-end learned policies, are insufficient in settings where tasks, environments, and human expectations are open-ended, long-tailed, and subject to adaptation over time. To address this gap, we propose modular safety guardrails, consisting of monitoring (evaluation) and intervention layers, as an architectural foundation for comprehensive safety across the autonomy stack. Beyond modularity, we highlight possible cross-layer co-design opportunities through representation alignment and conservatism allocation to enable faster, less conservative, and more effective safety enforcement. We call on the community to explore richer guardrail modules and principled co-design strategies to advance safe real-world physical AI deployment.

1. Introduction

Foundation models (FMs), large-scale networks pretrained for broad generalization, are rapidly becoming core components of modern robotic autonomy stacks (Firoozi et al., 2025; Siciliano et al., 2008). They are increasingly used for perception (Gadre et al., 2023), task planning (Zitkovich et al., 2023; Driess et al., 2023), and end-to-end visuomotor control (Kim et al., 2024; Black et al., 2025), enabling open-world semantic reasoning and cross-task generalization that push robots beyond controlled laboratory settings.

Embodiment fundamentally reshapes the safety problem (Kojima et al., 2025; Wu et al., 2024; Grislain et al., 2025). Classical robotics safety often assumes fixed and predefinable constraints (e.g., geometric collision bounds), whereas FM-enabled robots operate in open-ended environments where hazards are context-dependent and specifications evolve (Santos et al., 2025; Liu & Feng, 2024). Risk is further compounded by environmental uncertainty and FM stochasticity (Hafez et al., 2025; Dalrymple et al., 2024).

Moreover, interactions with humans impose safety requirements beyond physical feasibility, including semantic appropriateness, intent alignment, and adherence to social norms (Dragan et al., 2013; Tian & Oviatt, 2021; Brunke et al., 2025). These requirements induce diverse failure modes that cannot be addressed by a single mechanism (Liu et al., 2023): physical safety filters cannot infer that a "knife handoff" is contextually dangerous (Brunke et al., 2025), while semantic reasoners lack real-time enforcement needed to prevent collisions (Kojima et al., 2025). No monolithic safety mechanism reliably addresses all such failures.

A natural response is to learn a single end-to-end safety guardrail that jointly encodes physical, semantic, and intent constraints, but such monolithic solutions remain fragile in practice (Dawson et al., 2023). They are vulnerable to distribution shift as tasks, environments, and safety requirements evolve (Farid et al., 2022; Buysse et al., 2025), often necessitating additional frequently updated or externally imposed safety components (Ren et al., 2023; Peng et al., 2025). Moreover, real-world datasets rarely contain catastrophic safety failures, leaving the most critical modes underrepre-

† Equal co-lead contribution. ‡ Equal second-author contribution. * Equal senior advising. This work is not associated with Amazon. [1]Texas A&M University [2]Purdue University [3]Amazon [4]NVIDIA. Correspondence to: Yiwei Lyu <yiweilyu@tamu.edu>, Minghui Zheng <mhzheng@tamu.edu>, Yan Gu <yangu@purdue.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

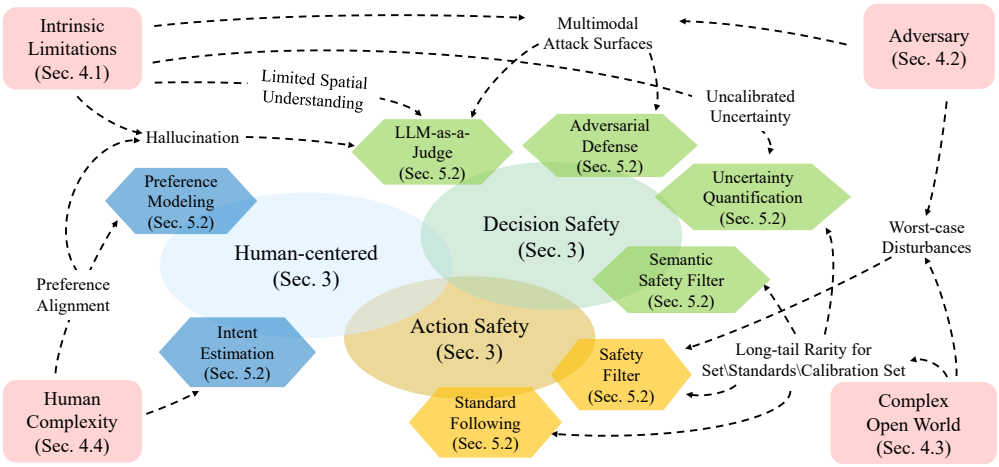

Figure 1. Overview of the safety definitions (Sec. 3), source of safety challenges (Sec. 4), and existing alternative methods (Sec. 5).

sented during training (Tölle et al., 2025). These limitations fundamentally constrain the robustness and longevity of purely end-to-end safety solutions.

In this position paper, we argue that architectural modularity is necessary for the safe and reliable deployment of FM-enabled robotic systems.

Core Claim

Under open-world deployment with long-tailed hazards and non-stationary human interaction, safety mechanisms that are either fully embedded within the autonomy model or confined to a single layer of the autonomy stack are structurally insufficient to ensure action, decision, and human-centered safety simultaneously.

In particular, our *external modularity* separates safety enforcement from FMs to prevent FM uncertainty from affecting safety, while *internal modularity* decomposes safety into specialized mechanisms targeting distinct failure modes.

We ground this claim with a three-dimensional taxonomy of FM-enabled robotics safety: action safety (physical feasibility and constraint compliance), decision safety (semantic and contextual appropriateness), and human-centered safety (conformance to human intent, norms, and expectations), as illustrated in Fig. 1. We argue that non-modular approaches are fundamentally brittle under real-world conditions, where safety requirements drift, are hard to specify a priori, and failures are rare but high impact.

To address these challenges, we propose a two-layer modular safety guardrail architecture (Fig. 2) with (i) a Monitoring and Evaluation Layer that assesses risk across the autonomy stack and (ii) an Intervention Layer that enforces safety through decision-level gating and action-level filtering. This modular design enables principled cross-layer co-design, such as representation alignment and conservatism allocation, allowing more precise and less conservative safety enforcement, while supporting independent verification, updateability, composability of heterogeneous mechanisms, and systematic coverage across all three safety dimensions.

This paper first reviews the integration of FM into robotic systems (Sec. 2), then formalizes the safety taxonomy (Sec. 3), analyzes safety challenges in FM-enabled robotics (Sec. 4), discusses alternative non-modular approaches (Sec. 5), and finally presents the modular safety guardrail architecture and its co-design opportunities (Sec. 6), where we explicitly characterize the architectural requirements for safe FM-robot deployment and introduce co-design principles that go beyond simple module stacking.

2. Roles of Foundation Models in Robotics

FM as a Perception Module. FMs are increasingly used as perceptual front ends in robotic systems, mapping raw sensory inputs such as images, depth, and language into high-level semantic representations (Ahn et al., 2022; Gorlo et al., 2025). For example, vision-language and multimodal FMs enable capabilities such as open-vocabulary object recognition (Liu et al., 2024a), scene understanding (Maggio & Carlone, 2025; Alama et al., 2025), and affordance prediction (Nasiriany et al., 2025), allowing robots to perceive previously unseen objects and environments without task-specific retraining (Gadre et al., 2023). By lifting perception from closed-set classification to semantic abstraction, FMs substantially expand a robot's operational scope, while also introducing new challenges related to uncertainty, grounding, and reliability in safety-critical settings (Ren et al., 2023; Huang et al., 2023; Kim et al., 2025).

FM as a Reasoning Module. Since the emergence of large language models (LLMs), a common integration of FMs in robotics is to use FMs as high-level semantic planners that interpret natural-language instructions and compose executable action sequences. Early systems grounded plan-

ning in predefined libraries of robot skills, translating user instructions into sequences of skill invocations. For example, given a skill set such as `move-to-<location>` or `grab-<object>`, an LLM can map an instruction like "bring me a bottle of water from the kitchen" into a structured plan. Representative studies include Code as Policies (Liang et al., 2022), ProgPrompt (Singh et al., 2023), PaLM-SayCan (Ahn et al., 2022), LLM-Planner (Song et al., 2023), and Alpamayo-R1 (Wang et al., 2025b).

FM as an Action Module. End-to-end vision-language-action (VLA) models extend this paradigm by adapting pretrained FM backbones to map images and language instructions directly to robot actions, unifying perception, grounding, and control in a single policy. RT-2 (Zitkovich et al., 2023) introduces an action-as-language approach by co-fine-tuning web-scale VLMs and representing actions as discrete autoregressive tokens. OpenVLA (Kim et al., 2024) similarly fine-tunes a Llama 2-based VLM into a 7B-parameter action generator trained on large-scale robot demonstrations. By contrast, the π-series (Black et al., 2025) preserves a pretrained VLM backbone (initialized from PaliGemma (Beyer et al., 2024)) and pairs it with a separate action expert for continuous control: $\pi_{0.5}$ (Black et al., 2025) extends π_0 (Black et al., 2024) via co-training on heterogeneous semantic supervision, while $\pi_{0.6}$ (Intelligence et al., 2025) scales the backbone (reported as 5B) and improves performance through on-robot experience. Gemini Robotics (Gemini Robotics Team et al., 2025) follows a similar end-to-end approach, fine-tuning a Gemini 2.0-based model to directly generate control commands.

3. Safety Definitions and Specifications for FM-enabled Robotics

This section introduces the *safety definitions* and associated *safety specifications* for FM-enabled robotic systems operating in human-centered, unstructured environments. We use *safety definitions* to describe conceptual categories of safety, and *safety specifications* to denote the explicit constraints that a robot must satisfy during operation.

Prior to the adoption of FMs, robotics safety primarily focused on *action safety*, which enforces low-level physical constraints. With FMs now integrated across the autonomy stack (Sec. 2), robots are increasingly deployed in unstructured, human-centered environments, inducing safety considerations that extend beyond physical execution. Thus, we organize safety into three complementary categories: *action safety*, *decision safety*, and *human-centered safety*.

❶ Action Safety. Action safety concerns maintaining a robot's physical execution within well-defined constraints, particularly in real-world environments (Hsu et al., 2023). Typical specifications include collision avoidance, adherence to joint limits, and safe force or impedance modulation during interaction. These safety specifications are relatively well understood and are largely addressed through model-based control-theoretic approaches (Hsu et al., 2023; Wabersich et al., 2023; Hewing et al., 2020).

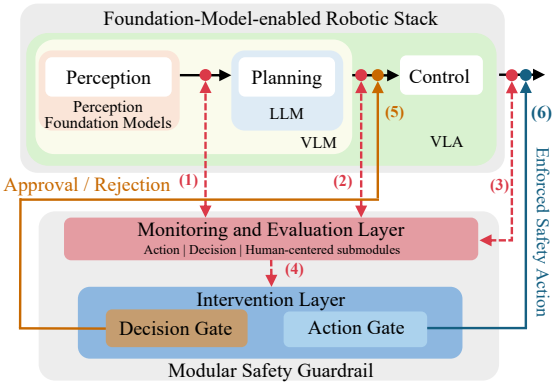

Figure 2. **Overview of one potential modular safety guardrail architecture.** It shows the architecture and information flow between the FM-enabled robotic stack **(top)** and the modular safety guardrail **(bottom)**. Arrows **1**–**3** denote information flow from perception, planning, and control to the Monitoring and Evaluation Layer, which uses dedicated submodules for each safety dimension to generate risk signals for downstream modules. Risk indicators from planning and control are sent to the Intervention Layer (Arrow **4**), consisting of a *Decision Gate* that screens plans and triggers replanning upon rejection (Arrow **5**) and an *Action Gate* that enforces physical safety constraints on control commands. The Action Gate may apply last-resort safety filters to ensure only physically safe actions are executed (Arrow **6**).

❷ Decision Safety. Decision safety generalizes action safety by incorporating semantic and contextual appropriateness in open-world settings. With LLMs integrated into robotic autonomy stacks, safety specifications must assess whether a robot's proposed behavior is appropriate with respect to open-domain knowledge and task semantics (Nakamura et al., 2025). For example, constraints such as *"a soft toy must not be placed on a hot stove"* (Gemini Robotics Team et al., 2025) and *"not pouring coffee too fast"* (Nakamura et al., 2025) illustrate decision-level safety requirements that cannot be captured by low-level constraints alone.

❸ Human-centered Safety. Human-centered safety concerns whether robot behavior is perceived by humans as predictable, understandable, and trustworthy over sustained interaction, such that physical, cognitive, and social risks remain acceptable (Hancock et al., 2011; Kress-Gazit et al., 2021). Even when action- and decision-level constraints are satisfied, misaligned expectations, non-stationary human adaptation, and miscalibrated trust can cause behavior to be perceived as unsafe (Sagheb et al., 2025; Peng et al., 2025).

Intersections across dimensions. These three dimensions are not mutually exclusive. Many real-world failures span

multiple dimensions simultaneously. For example, handing scissors to a child may pass action-level feasibility checks while violating decision-level appropriateness and human-centered norms; an action-only filter would approve such a handoff, while a semantic checker alone would lack real-time physical enforcement. Addressing such overlaps motivates the cross-layer and cross-module co-design described in Sec. 6.3.

4. Safety Challenges in FM-enabled Robotics

This section analyzes the primary safety challenges faced by FM-enabled robotic systems and their implications for the safety definitions introduced in Sec. 3, as illustrated in Fig. 1. We group these challenges by source, highlighting why modular runtime safety mechanisms are required beyond end-to-end learning alone.

4.1. Intrinsic FM and Robot Limitations

Many safety failures in FM-enabled robotics stem from intrinsic limitations of FMs as perceptual and reasoning components, as well as robots as physical systems. FM-enabled robotics inherits safety concerns already studied in perception, VLM, and LLM systems. However, embodiment introduces an additional safety burden: perception or reasoning errors can propagate through planning and control into unsafe physical behavior. In particular, epistemic unreliability in perception and reasoning primarily threatens *decision safety*, while hard physical constraints and modeling mismatch threaten *action safety*.

Epistemic unreliability under distribution shift. Despite rapid progress, VLM/VLA systems remain brittle under distribution shift, exhibiting hallucination, weak spatial grounding, and poorly calibrated uncertainty (Liu et al., 2024b; Chakraborty et al., 2025b; Chen et al., 2024a). For generative models, reliability failures extend beyond input out-of-distribution (OOD) detection to the trustworthiness of generated outputs, which is not well captured by conventional uncertainty measures (Ovadia et al., 2019; Xu & Ding, 2025). These failures can propagate to downstream decisions and induce unsafe behaviors.

Hard physical constraints and long-tail failures. Action safety is constrained by hardware limits, unmodeled dynamics, and environment-imposed constraints (Hsu et al., 2023). Safety-critical physical failures are inherently rare and dominated by worst-case interactions (Brunke et al., 2022; Kojima et al., 2025; He et al., 2025), creating a mismatch between statistical learning objectives and hard constraint satisfaction. Explicit low-level enforcement therefore remains necessary for open-world deployment.

4.2. Adversarial and Worst-case External Factors

A second source of risks comes from adversarial manipulation and worst-case external factors, which threaten *decision safety* via perception- and language-level attacks and *action safety* via disturbances that violate nominal assumptions. Multimodal FM-enabled control pipelines expose expanded attack surfaces (e.g., prompt- and perception-level manipulations) that can bypass intended safeguards (Jones et al., 2025; Robey et al., 2025; Xing et al., 2025), while real-world operation entails disturbances and unexpected contacts that are difficult to model exhaustively. This motivates runtime safety mechanisms that are decoupled from FM-based decision modules, such as certified control or safety-filtering layers that enforce physical constraints independently of high-level reasoning (Hsu et al., 2023; Ames et al., 2019).

4.3. Open-world Deployment

FM-enabled robots increasingly operate in open-world settings where tasks, constraints, and contexts are compositional and effectively unbounded (Firoozi et al., 2025). This shift makes exhaustive specification, testing, and offline validation infeasible. Such deployment induces task-space explosion and long-tail safety risks, where rare but critical failure modes dominate overall risk (He et al., 2025; Angelopoulos et al., 2023). Ensuring safety therefore requires lifecycle-level strategies that span prevention, runtime monitoring and mitigation, and post-incident recovery, rather than relying solely on pre-deployment training or static rules (Kojima et al., 2025; Tan et al., 2025).

4.4. Human Complexity and Mutual Adaptation

Human-centered safety challenges arise from the interactive, subjective, and evolving nature of human behavior and expectations, primarily defining *human-centered safety* while also feeding back into *decision safety* through ambiguous intent and dynamic preference shifts. Language-enabled interaction introduces ambiguity in user intent (Ren et al., 2023; Santos et al., 2025; Mehta et al., 2024), and human-perceived safety is inherently subjective, context-dependent, and personalized, making it difficult to encode as fixed constraints or universal objectives (Santos et al., 2025; Mehta et al., 2024). Even when action- and decision-level constraints are satisfied, mutual adaptation, shifting trust, and evolving social norms can cause robot behavior to be perceived as unsafe (Tan et al., 2025), rendering static thresholds and uncertainty-only formulations insufficient for sustained human–robot interaction.

5. Alternative Views

This section organizes existing safety literature around alternative views on *where* safety should be placed within

FM-enabled robotic systems. These views include embedding safety considerations within end-to-end models, as well as applying isolated external add-ons adapted from perception, LLM, and VLM safety studies, such as alignment, uncertainty estimation, semantic filtering, and preference modeling. Each view addresses a subset of the safety challenges in Sec. 4, but none provides comprehensive protection across action, decision, and human-centered risks in open-world deployment. This analysis motivates the need for the modular guardrail architecture introduced in Sec. 6.

5.1. View A: Safety Should be Embedded in the Model

One alternative view argues that safety should be internalized directly within the FM or base policy, such that safe behavior emerges by construction rather than external constraints. This view is supported by post-training alignment methods, such as instruction-tuning and reinforcement learning from human feedback (Ouyang et al., 2022), which can substantially shift model behavior toward user intent and reduce undesirable outputs. Related efforts, such as principle-based alignment (e.g., Constitutional AI (Bai et al., 2022; Gemini Robotics Team et al., 2025)) and recent safety pre-training proposals (Maini et al., 2025), further advocate treating safety as a first-class training objective so that internal representations and decision rules are safety-aware. **What it cannot cover?** Embedding safety within the model does not establish a non-bypassable runtime boundary between high-level decision-making and physical execution. Even well-aligned models remain vulnerable to distribution shift, novel hazards, and unforeseen open-world interactions. While increasingly capable end-to-end systems may reduce reliance on modular safeguards in controlled settings, we argue that externally enforceable modular guardrails remain necessary for open-world deployment with evolving constraints and long-tail failures. When failures occur, there is no external mechanism to prevent unsafe actions from being executed. As a result, model-internal safety alone cannot guarantee action-level safety at deployment time.

Recent VLA systems also suggest that modular adaptation can remain useful within otherwise end-to-end pipelines. For example, RL Token (Xu et al., 2026) adapts behavior through lightweight residual modules rather than fully retraining the entire VLA. Similarly, adapting modular safety guardrails to deployment-specific requirements may often be more practical than fully fine-tuning large foundation models for every downstream safety domain.

5.2. View B: Safety Should be Provided by External Add-ons (Single-perspective or Narrow Subset)

A second view places safety outside the FM, implemented through external modules that monitor, filter, or constrain system behavior at runtime. Most existing approaches adopt a single perspective, addressing only one safety dimension or a narrow subset rather than providing integrated coverage across all three safety dimensions.

View B1: Action-level add-ons. Action-level approaches include safety standards (ANSI/RIA, 2012; ISO, 2011; Jacobs & Virk, 2014) and control-theoretic safety filters (Margellos & Lygeros, 2011; Fisac et al., 2019; Ames et al., 2019; Bastani, 2021; Wabersich & Zeilinger, 2018) that intervene during execution to enforce physical constraints. These methods are effective for ensuring collision avoidance, constraint satisfaction, and physical feasibility. **What it cannot cover?** Action-level mechanisms cannot reason about semantic hazards, task-level mistakes, or harmful intent, nor can they capture human-contextual risks. They also rely on predefined unsafe sets and sufficiently accurate models of system dynamics, assumptions that often break down for FM-enabled robots operating in unstructured environments (Hsu et al., 2023; Bajcsy & Fisac, 2024).

View B2: Decision-level add-ons. Decision-level approaches attempt to detect unsafe plans or commands before execution, using learned world models (Nakamura et al., 2025; Seo et al., 2025; Agrawal et al., 2025), LLM-based judges (Gu et al., 2024; Duan et al., 2024; Yang et al., 2025; Gao et al., 2025; Khan et al., 2025; Ravichandran et al., 2025; Jindal et al., 2025) or uncertainty estimation techniques (Ren et al., 2023; Liang et al., 2024; Sun et al., 2024b;a; Wang et al., 2025a; Karli et al., 2025). These methods directly target decision-level safety by filtering or modifying high-level outputs. **What it cannot cover?** They cannot guarantee physical safety during execution (Bajcsy & Fisac, 2024), are vulnerable to hallucinations and adversarial manipulation when LLMs are involved (Chen et al., 2024b; Gao et al., 2024; Xing et al., 2024; Xu et al., 2025; Jones et al., 2025; Lechner et al., 2023; Everett et al., 2021), and often assume access to complete or correct safety specifications. Additionally, their performance degrades significantly under distribution shift, limiting reliability in real-world deployment (He et al., 2025).

View B3: Human-centered add-ons. Human-centered approaches rely on intent inference, preference learning, trust modeling, and feedback-based adaptation to align robot behavior with human expectations (Peng et al., 2025; Dixit et al., 2023; Chakraborty et al., 2025a; Salzmann et al., 2020; Chen et al., 2025; Pandya et al., 2025). These methods improve interaction quality and personalization. **What it cannot cover?** They cannot provide hard safety guarantees or enforcement to prevent unsafe actions when misalignment occurs, as human intent is inherently ambiguous, preferences change over time, multi-human environments introduce conflicting constraints (Shi et al., 2025).

Key Takeaway

Neither model-internal safety nor single-perspective external add-ons are sufficient in isolation. Internal alignment improves typical behavior, but cannot replace non-bypassable enforcement at execution time, while external add-ons usually address only one safety dimension. Ensuring safety in FM-enabled robotic systems, therefore, requires a modular safety guardrail architecture that provides integrated coverage across action-level, decision-level, and human-centered risks, as developed in Sec. 6. These insufficiencies motivate the architectural requirements (Sec. 6.1) and co-design principles (Sec. 6.3); stacking such individually insufficient mechanisms would inherit the same gaps.

6. Modular Safety Guardrails

6.1. Definition and Design Principles

We argue that architectural modularity is necessary for the safe and reliable deployment of FM-enabled robotic systems. As discussed in Sec. 4, such systems face non-stationary and hard-to-predefine safety specifications, compounded uncertainty from both the environment and FMs, and rare but catastrophic safety failures. Addressing these challenges requires safety mechanisms that are (i) **independently verifiable and updateable** as requirements evolve, (ii) **composable** across complementary failure modes spanning physical, semantic, and human-interaction contexts, and (iii) **non-bypassable at execution time**.

We capture these requirements with a *modular safety guardrail*: a safety layer decoupled from upstream autonomy components that supports independent updates, composition of heterogeneous mechanisms, and enforceable closed-loop intervention. Here, we use *guardrails* as an umbrella term for modular safety mechanisms consisting of (i) *monitoring* components that evaluate safety conditions and (ii) *intervention* components that constrain or override actions when violations are detected, as introduced in Sec. 6.2.

Definition 1 (Modular safety guardrail). A modular safety guardrail is a non-bypassable, execution-time safety architecture that mediates all execution-relevant proposals (e.g., perceptions, plans, and commands) from upstream autonomy components before they reach the robot's physical execution layer. Operating in the closed-loop robot-environment-human system, it monitors risk and intervenes at runtime to prevent unsafe behavior across physical, semantic, and interaction-level safety dimensions.

A modular safety guardrail is characterized by two forms of modularity: *external* and *internal*.

External modularity. An external guardrail is designed to be operationally independent of the upstream FMs it monitors. It should not share parameters, training objectives, or optimization procedures with those FMs, and should minimize statistical coupling (e.g., shared pretraining corpora or safety-annotation data) that could induce correlated errors. This independence enables the guardrail to function as an auditable, verifiable safety authority rather than inheriting the same uncertainty and failure modes as the guarded FM.

Internal modularity. Internal modularity decomposes the guardrail into specialized submodules with explicit interfaces and non-overlapping safety responsibilities, each targeting distinct failure modes and time-scale requirements (e.g., perception trust assessment, decision-level semantic or intent screening, and action-level constraint enforcement). Submodules may consume different autonomy-stack signals, yet remain independently testable, replaceable, and updateable without retraining the entire guardrail. This structure limits cross-dimension error propagation and supports robust enforcement under evolving specifications and real-time constraints.

6.2. Modular Safety Guardrail Architecture

The modular safety guardrail decomposes safety enforcement into two functionally distinct layers (Fig. 2, bottom): a *Monitoring and Evaluation Layer* that **assesses risk** across autonomy-stack outputs, and an *Intervention Layer* that **enforces safety** through two complementary submodules: a *decision gate* and an *action gate*. This two-layer architecture instantiates the internal modularity principle of Sec. 6.1, while external modularity is preserved through the guardrail's operational independence from the upstream FM stack. The further subdivision of the Intervention Layer into decision and action gates represents a fine-grained decomposition of the same principle. Together, these layers address the complementary failure modes across the three safety dimensions identified in Sec. 3.

6.2.1. Monitoring & Evaluation Layer

The Monitoring and Evaluation Layer performs independent risk assessment across the autonomy stack, aiming to detect potential safety violations before they propagate downstream. It observes execution-relevant outputs from perception, planning, and control through channels decoupled from the primary autonomy pipeline. These channels may include auxiliary sensors for cross-validating perception (Antonante et al., 2023a;b), FM-based evaluators (e.g., critic or red-teaming agents) for assessing high-level plans (Elhafsi et al., 2023; Sinha et al., 2024), and risk-aware control-theoretic monitors for detecting physical constraint violations (Frank et al., 2024; Lyu et al., 2023). To help reduce shared failure modes between the system being evaluated and the evaluator, this layer interfaces with explicit, execution-relevant

representations (e.g., regions of interest, candidate plans and trajectories, and control commands) and does not rely on internal latent embeddings unless they are explicitly exposed.

The layer can produce risk signals spanning all three safety dimensions (through dedicated submodules). Not all safety properties are equally monitorable or formally specifiable. Low-level physical constraints, such as joint limits, collision margins, or velocity bounds, are often amenable to runtime monitoring and enforcement. In contrast, higher-level semantic and human-centered safety properties may only admit approximate, probabilistic, or human-in-the-loop evaluation, and exact online verification may be computationally intractable in real-time embodied settings. The proposed modular architecture does not resolve these formal limitations, but instead decomposes safety into subproblems with different monitoring and computational requirements. For action safety, it flags violations of joint limits, collision margins, or dynamic feasibility. For decision safety, it checks semantic consistency and contextual appropriateness, including FM-specific issues such as hallucination or adversarial vulnerability. For human-centered safety, it monitors predictability, preference alignment, and trust-related indicators. These signals are passed independently to the Intervention Layer, which executes non-bypassable mitigation and coordinates enforcement across dimensions via co-design (Sec. 6.3).

6.2.2. Intervention Layer

Because robots are physically embodied, recognizing risk is not enough: unsafe proposals must be blocked or modified before reaching the actuators. The *Intervention Layer* provides this non-bypassable, execution-time authority by applying concrete mitigations when monitored risk exceeds acceptable thresholds, through two complementary mechanisms: a planning-level *decision gate* and an execution-time *action gate*.

This two-gate design reflects that safety risks arise at different semantic levels and time scales. High-level failures (e.g., misinterpreted intent, unsafe semantics, and norm violations) should be intercepted before execution commits the robot to an inappropriate course of action. Conversely, even semantically appropriate plans may become unsafe under disturbances, tracking error, state-estimation drift, or unmodeled contacts, requiring an action gate as the last line of defense. Separating these roles localizes responsibility and supports principled conservatism allocation: reject plans only when necessary, and otherwise enforce safety through minimal execution-time modification.

Decision Gate. The decision gate operates at the planning level, screening candidate plans using aggregated risk signals from the Monitoring and Evaluation Layer (Fig. 2, solid brown arrow (5)). It targets violations of decision safety and human-centered safety in FM-generated plans. We design the gate as a filter that is explicitly decoupled from plan generation, preserving a clear safety authority boundary between proposing actions and approving them for execution. When risk exceeds acceptable thresholds, the gate blocks the plan and triggers replanning or user clarification. By filtering semantically or socially unsafe plans upstream, the decision gate prevents the action gate from compensating for fundamentally inappropriate intent, reducing unnecessary stops and overly conservative execution.

Action Gate. The action gate operates at the execution level, enforcing physical safety by constraining or modifying low-level control commands before they are applied to the robot (Fig. 2, solid navy arrow (6)). It enforces action safety via shielding, trajectory adjustment, or projection onto safe sets, with constraints parameterized by monitoring outputs such as uncertainty or trust. Unlike the decision gate, which reasons over plan content, the action gate provides non-bypassable physical enforcement regardless of how plans are generated. Architecturally, it is the last line of defense: even after a plan is approved, it ensures executed commands remain within acceptable physical bounds under disturbances and residual upstream failures.

Safety Assurance. The safety assurances of the proposed architecture come from enforcing a clear safety authority boundary rather than from assuming any single model is correct. Concretely, (i) non-bypassability ensures all execution-relevant proposals pass through the intervention layer before reaching the actuators; (ii) external modularity/operational independence enables independent auditing and verification; and (iii) internal modularity allows heterogeneous safety mechanisms to be verified, updated, and composed without retraining the full stack. Together, these properties provide enforceable runtime safety envelopes (via the action gate) and upstream rejection of semantically or socially unsafe plans (via the decision gate), yielding systematic coverage across action, decision, and human-centered safety.

6.3. Co-Design Opportunities Enabled by Modular Guardrails and Deployment Examples

The modular safety guardrail architecture provides an enforceable foundation for comprehensive safety across action, decision, and human-centered dimensions. Beyond this foundation, the architecture also exposes a new space for co-design that can make safety enforcement faster, less conservative, and more graceful in practice. We view possible co-design systematically along two axes: (i) between layers (Monitoring and Evaluation and Intervention Layers) and (ii) within a layer (coordination among modules inside the Intervention Layer). We highlight two key opportunities.

Representation alignment: Co-design between layers. Representation alignment concerns the interaction between

layers, i.e., how safety-relevant information produced by the Monitoring and Evaluation Layer is represented so that the Intervention Layer can make more informed interventions. The core principle is representation compatibility: monitoring outputs should preserve uncertainty and risk structure that downstream enforcement primitives can act on directly, rather than collapsing them into lossy scalar scores. For example, expressing perception uncertainty as a spatially grounded set (e.g., ellipsoidal pose uncertainty, occupancy tubes, and workspace-indexed risk fields) allows the action gate to tighten constraints directionally or locally where risk is concentrated, instead of applying uniform conservatism everywhere. In this way, co-design at the layer interface turns "risk assessment" into actionable, enforcement-ready inputs that support precise intervention.

Conservatism allocation: Co-design between modules. Conservatism allocation concerns interaction among modules inside the Intervention Layer, especially between the decision gate and the action gate. If each module applies its strictest criterion independently, conservatism stacks, often producing infeasible behavior (premature plan rejection, excessive projection, or unnecessary stoppages). Co-design enables coordinated strictness: the decision gate can approve plans conditionally on whether the action gate can enforce the induced margins in real time, and can escalate to rejection or human clarification only when the action gate reports infeasibility. This shifts conservatism to the module that can enforce it most precisely, preserving task progress while maintaining safety.

Together, these two strategies clarify how the proposed architecture enables more than modular enforcement: co-design across layers improves what information is enforced, while co-design within layers improves how enforcement authority is exercised without compounding conservatism.

Next, we provide three[1] deployment examples to illustrate how the proposed modular safety guardrail interfaces with different FM configurations. Each example serves as an existence proof of a concrete failure mode that cannot be addressed by model-internal or single-perspective safety alone, and demonstrates how cross-layer or cross-module co-design enables effective resolution in practice.

Example A: Language-Instructed Household Robot (LLM/VLM-based Planner). A household robot follows natural-language instructions (e.g., "clean up the living room") using an LLM planner to generate skill sequences (Singh et al., 2023; Ahn et al., 2022). Such planners can produce unsafe or ill-posed plans due to semantic misinterpretation, ambiguous intent, or hallucinated affordances (e.g., discarding items to be preserved, placing a hot pan on wood, or assuming a drawer is open). The monitoring layer checks semantic consistency via LLM-as-a-Judge (Elhafsi et al., 2023; Sinha et al., 2024), quantifies uncertainty with conformal prediction (Ren et al., 2023), verifies constitutional rules (Sermanet et al., 2025; Jindal et al., 2025), and evaluates human-centered risks (predictability and intent alignment). The decision gate rejects or requests clarification when risk or uncertainty is high, while the action gate enforces physical constraints through trajectory projection onto safe sets (Fisac et al., 2019).

A co-design example of representation alignment and conservatism allocation: While carrying a hot pan from the stove to the counter, the robot observes a child entering and moving toward its planned path. Without co-design, the decision gate may apply a fixed semantic rule (e.g., "hot object near child") and reject the plan, freezing the robot mid-motion while holding a hazard. With co-design, the monitoring layer converts this semantic hazard into the same object the action gate can enforce: a time-varying keep-out region (or occupancy tube) for the child, with a radius scaled by thermal risk rather than collision-only risk (representation alignment). The decision gate then allocates conservatism by approving continuation whenever the action gate can certify feasibility of maintaining separation under these means (conservatism allocation), and escalating only when it cannot. In execution, the action gate enlarges clearance (e.g., 1.0 m for thermal hazards vs. 0.3 m for collision-only), reduces speed, and reroutes to a farther placement location, allowing the robot to safely complete the placement instead of defaulting to a brittle stop.

6.4. Scope and Limitation

The modular safety guardrail is not a universal solution to safety in FM-enabled robotics. It is not intended to make the FMs intrinsically safer or to guarantee globally optimal decisions. Rather, it provides an enforceable runtime layer that detects unreliable outputs and mitigates their consequences, complementing offline retraining or fine-tuning, which cannot be relied on during online execution. We do not claim that modular architectures will always outperform alternative safety approaches (e.g., end-to-end safety), but rather that externally enforceable runtime safeguards remain necessary for practical open-world deployment. In this sense, the guardrail improves deployability by expanding safety enforcement beyond physical constraints, enabling intervention on contextually inappropriate high-level decisions while retaining action-level gating as the last line of defense against physical harm.

Another motivation for modularity comes from heterogeneous complexity and verification requirements across safety dimensions. Some safety properties may admit formal verification or tractable runtime checking, while oth-

[1]Due to space considerations, one representative example is included in the main text, with two additional ones in Appendix A.

ers involving open-world semantics, human intent, or long-horizon reasoning may be computationally intractable or difficult to fully specify in a decidable form. The proposed modular architecture allows different safety mechanisms to operate under different computational and formal assumptions, enabling practical real-time enforcement without requiring a single globally verifiable end-to-end safety model.

Another practical limitation is that the effectiveness of the guardrail depends on the independence of its safety signal from the FM-based autonomy stack it monitors. When the guardrail itself relies on FMs, such "independence" may be hard to ensure in practice because models from different sources can share training data, design choices, or evaluation pipelines (e.g., LLMs/VLMs like Qwen (Bai et al., 2025), Llama (Touvron et al., 2023), Gemini (Comanici et al., 2025), GPT (Achiam et al., 2023) or VLAs like OpenVLA, GR00T (Bjorck et al., 2025), Gemini Robotics). This overlap can induce correlated failure modes, weakening FM-based guardrails as independent checks. These challenges motivate clearer notions of relative or operational independence, including metrics that quantify statistical dependence or shared failure risk, to characterize when FM-based guardrails provide meaningful safety benefits.

A further practical consideration is runtime execution. FM-based guardrails introduce computational overhead, and we view this as a necessary cost of test-time scaling for safety (Kwok et al., 2025; Wu et al., 2025), where additional computation is allocated to safety-critical monitoring and intervention. Although modular architectures can mitigate latency through parallel safety intervention between the decision and action gates, the decision gate may still incur delays from large-scale FMs needed to reason about complex causal and contextual relationships. Practical deployment therefore requires cross-module co-design that accounts for gate asynchrony, including fallback mechanisms that allow the action gate to apply conservative adjustments while the decision gate completes its reasoning.

7. Call to Action & Conclusion

We argue that safety for FM-enabled robotics must be approached as a system property, requiring explicit mechanisms that jointly address action, decision, and human-centered risks. We advocate modular safety guardrails as a practical architectural foundation to decouple safety authority from any single model, support independent auditing and updates, and enable robust deployment across tasks and platforms. Our goal is not to propose a complete safety solution, but to identify architectural conditions that are necessary for any safety mechanism to scale to open-world FM-enabled robotic systems. While grounded in robotics, we view the safety challenges and architecture described in this paper as broadly applicable to embodied, agentic AI systems that couple FM reasoning with real-world action under uncertainty. We invite the community to help turn this architectural position into a practical, shared safety stack along two complementary directions.

First, we call for a broader ecosystem of composable guardrail modules. Beyond basic monitoring and intervention components, there is a large design space for modules that target specific failure modes, uncertainty sources, and human interaction contexts. To enable reuse across platforms and tasks, such modules should expose clear interfaces and semantics. Progress here also depends on evaluation: we encourage the development of comprehensive benchmarks that explicitly test action, decision, and human-centered safety, including rare but catastrophic scenarios that are systematically underrepresented in today's datasets.

Second, we call for principled co-design within the guardrail architecture. Co-design goes beyond assembling modules; it requires specifying what safety-relevant information is produced (e.g., risk, uncertainty, and constraint structure), how it is represented, and how it flows bidirectionally across the autonomy stack. Such co-design allows upstream components to learn to avoid repeatedly proposing actions that downstream gates will reject or heavily modify. With appropriate co-design, safety can be faster to enforce, less conservative in practice, and more robust to deploy, by allocating conservatism where it is needed based on a systematic view instead of accumulating it everywhere.

We hope this position paper serves as a starting point for a shared research agenda: to standardize interfaces, expand modular guardrail capabilities, and develop co-design principles that yield composable, updateable, and deployment-ready comprehensive safety mechanisms for FM-enabled robotic systems and related embodied AI platforms.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, S., Seo, J., Nakamura, K., Tian, R., and Bajcsy, A. Anysafe: Adapting latent safety filters at runtime via safety constraint parameterization in the latent space. *arXiv preprint arXiv:2509.19555*, 2025.

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.

Alama, O., Jariwala, D., Bhattacharya, A., Kim, S., Wang, W., and Scherer, S. Radseg: Unleashing parameter

and compute efficient zero-shot open-vocabulary segmentation using agglomerative models. *arXiv preprint arXiv:2511.19704*, 2025.

Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. Control barrier functions: Theory and applications. In *European control conference*, pp. 3420–3431, 2019.

Angelopoulos, A. N., Bates, S., et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.

ANSI/RIA. Industrial robots and robot systems – safety requirements. Standard ANSI/RIA R15.06-2012, Robotic Industries Association, 2012.

Antonante, P., Nilsen, H. G., and Carlone, L. Monitoring of perception systems: Deterministic, probabilistic, and learning-based fault detection and identification. *Artificial Intelligence*, 325:103998, 2023a.

Antonante, P., Veer, S., Leung, K., Weng, X., Carlone, L., and Pavone, M. Task-aware risk estimation of perception failures for autonomous vehicles. *arXiv preprint arXiv:2305.01870*, 2023b.

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

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Bajcsy, A. and Fisac, J. F. Human-ai safety: A descendant of generative ai and control systems safety. *arXiv preprint arXiv:2405.09794*, 2024.

Bastani, O. Safe reinforcement learning with nonlinear dynamics via model predictive shielding. In *American control conference*, pp. 3488–3494, 2021.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Galliker, M. Y., Ghosh, D., Groom, L., Hausman, K., brian ichter, Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A. Z., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., and Zhilinsky, U. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Conference on Robot Learning*, 2025.

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Brunke, L., Zhang, Y., Römer, R., Naimer, J., Staykov, N., Zhou, S., and Schoellig, A. P. Semantically safe robot manipulation: From semantic scene understanding to motion safeguards. *IEEE Robotics and Automation Letters*, 2025.

Buysse, L., Habli, I., Vanoost, D., and Pissoort, D. Safe autonomous systems in a changing world: Operationalising dynamic safety cases. *Safety Science*, 191:106965, 2025.

Chakraborty, K., Feng, Z., Veer, S., Sharma, A., Ding, W., Topan, S., Ivanovic, B., Pavone, M., and Bansal, S. Safety evaluation of motion plans using trajectory predictors as forward reachable set estimators. *arXiv preprint arXiv:2507.22389*, 2025a.

Chakraborty, N., Ornik, M., and Driggs-Campbell, K. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*, 57(7):1–35, 2025b.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., and Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Chen, H., Li, S., Fan, J., Duan, A., Yang, C., Navarro-Alarcon, D., and Zheng, P. Human-in-the-loop robot learning for smart manufacturing: A human-centric perspective. *IEEE Transactions on Automation Science and Engineering*, 2025.

Chen, M., Tu, J., Qi, C., Dang, Y., Zhou, F., Wei, W., and Yin, J. Towards physically realizable adversarial attacks in embodied vision navigation. *arXiv preprint arXiv:2409.10071*, 2024b.

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

D., Rosen, E., 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.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

Dawson, C., Gao, S., and Fan, C. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 39(3):1749–1767, 2023.

Dixit, A., Lindemann, L., Wei, S. X., Cleaveland, M., Pappas, G. J., and Burdick, J. W. Adaptive conformal prediction for motion planning among dynamic agents. In *Learning for Dynamics and Control Conference*, pp. 300–314, 2023.

Dragan, A. D., Lee, K. C., and Srinivasa, S. S. Legibility and predictability of robot motion. In *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 301–308, 2013.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al. Palm-e: An embodied multimodal language model. 2023.

Duan, J., Pumacay, W., Kumar, N., Wang, Y. R., Tian, S., Yuan, W., Krishna, R., Fox, D., Mandlekar, A., and Guo, Y. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024.

Elhafsi, A., Sinha, R., Agia, C., Schmerling, E., Nesnas, I. A., and Pavone, M. Semantic anomaly detection with large language models. *Autonomous Robots*, 47(8):1035–1055, 2023.

Everett, M., Lütjens, B., and How, J. P. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4184–4198, 2021.

Farid, A., Veer, S., and Majumdar, A. Task-driven out-of-distribution detection with statistical guarantees for robot learning. In *Conference on Robot Learning*, pp. 970–980, 2022.

Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K., et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

Fisac, J. F., Lugovoy, N. F., Rubies-Royo, V., Ghosh, S., and Tomlin, C. J. Bridging hamilton-jacobi safety analysis and reinforcement learning. In *IEEE International Conference on Robotics and Automation*, pp. 8550–8556, 2019.

Frank, D.-A., Chrysochou, P., Mitkidis, P., Otterbring, T., and Ariely, D. Navigating uncertainty: Exploring consumer acceptance of artificial intelligence under self-threats and high-stakes decisions. *Technology in Society*, 79:102732, 2024.

Gadre, S. Y., Wortsman, M., Ilharco, G., Schmidt, L., and Song, S. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23171–23181, 2023.

Gao, X., Lin, Q., Luo, C., Xie, W., Shen, L., Kusumam, K., and Song, S. Scale-free and task-generic attack: Generating photo-realistic adversarial patterns with patch quilting generator. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2985–2989, 2024.

Gao, X., Lin, T.-H., Song, R., Wu, Y., Huang, K.-R., Jin, Z., Lin, F., Liu, S., and Tu, Z. Safecoop: Unravelling full stack safety in agentic collaborative driving. *arXiv preprint arXiv:2510.18123*, 2025.

Gemini Robotics Team, Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Gorlo, N., Schmid, L., and Carlone, L. Describe anything anywhere at any moment. *arXiv preprint arXiv:2512.00565*, 2025.

Grislain, C., Rahimi, H., Sigaud, O., and Chetouani, M. I-failsense: Towards general robotic failure detection with vision-language models. *arXiv preprint arXiv:2509.16072*, 2025.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Hafez, A., Akhormeh, A. N., Hegazy, A., and Alanwar, A. Safe llm-controlled robots with formal guarantees via reachability analysis. *arXiv preprint arXiv:2503.03911*, 2025.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

He, L., Jia, Q.-S., Li, A., Sang, H., Wang, L., Lu, J., Zhang, T., Zhou, J., Zhang, Y., Wang, Y., et al. Towards provable probabilistic safety for scalable embodied ai systems. *arXiv preprint arXiv:2506.05171*, 2025.

Hewing, L., Wabersich, K. P., Menner, M., and Zeilinger, M. N. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):269–296, 2020.

Hsu, K.-C., Hu, H., and Fisac, J. F. The safety filter: A unified view of safety-critical control in autonomous systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2023.

Huang, C., Mees, O., Zeng, A., and Burgard, W. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation*, pp. 10608–10615, 2023.

Intelligence, P., Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., et al. $\pi_{0.6}$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025.

ISO. Robots and robotic devices – safety requirements for industrial robots – part 1: Robots. Standard ISO 10218-1:2011, ISO, Geneva, Switzerland, 2011.

Jacobs, T. and Virk, G. S. Iso 13482-the new safety standard for personal care robots. In *International Symposium on Robotics*, pp. 1–6, 2014.

Jindal, A., Kalashnikov, D., Chang, O., Garikapati, D., Majumdar, A., Sermanet, P., and Sindhwani, V. Can ai perceive physical danger and intervene? *arXiv preprint arXiv:2509.21651*, 2025.

Jones, E. K., Robey, A., Zou, A., Ravichandran, Z., Pappas, G. J., Hassani, H., Fredrikson, M., and Kolter, J. Z. Adversarial attacks on robotic vision language action models. *arXiv preprint arXiv:2506.03350*, 2025.

Karli, U. B., Kurumisawa, T., and Fitzgerald, T. Ask before you act: Token-level uncertainty for intervention in vision-language-action models. In *Second Workshop on Out-of-Distribution Generalization in Robotics at RSS*, 2025.

Khan, A. A., Andrev, M., Murtaza, M. A., Aguilera, S., Zhang, R., Ding, J., Hutchinson, S., and Anwar, A. Safety aware task planning via large language models in robotics. *arXiv preprint arXiv:2503.15707*, 2025.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Kim, S., Alama, O., Kurdydyk, D., Keller, J., Keetha, N., Wang, W., Bisk, Y., and Scherer, S. Raven: Resilient aerial navigation via open-set semantic memory and behavior adaptation. *arXiv preprint arXiv:2509.23563*, 2025.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Kojima, T., Zhu, Y., Iwasawa, Y., Kitamura, T., Yan, G., Morikuni, S., Takanami, R., Solano, A., Matsushima, T., Murakami, A., et al. A comprehensive survey on physical risk control in the era of foundation model-enabled robotics. *arXiv preprint arXiv:2505.12583*, 2025.

Kress-Gazit, H., Eder, K., Hoffman, G., Admoni, H., Argall, B., Ehlers, R., Heckman, C., Jansen, N., Knepper, R., Křetínskỳ, J., et al. Formalizing and guaranteeing human-robot interaction. *Communications of the ACM*, 64(9): 78–84, 2021.

Kwok, J., Agia, C., Sinha, R., Foutter, M., Li, S., Stoica, I., Mirhoseini, A., and Pavone, M. Robomonkey: Scaling test-time sampling and verification for vision-language-action models. *arXiv preprint arXiv:2506.17811*, 2025.

Lechner, M., Amini, A., Rus, D., and Henzinger, T. A. Revisiting the adversarial robustness-accuracy tradeoff in robot learning. *IEEE Robotics and Automation Letters*, 8 (3):1595–1602, 2023.

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. Code as policies: Language model programs for embodied control. *arXiv preprint arXiv:2209.07753*, 2022.

Liang, K., Zhang, Z., and Fisac, J. F. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. *Advances in Neural Information Processing Systems*, 37: 71998–72031, 2024.

Liu, F., Fang, K., Abbeel, P., and Levine, S. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA*, 2024a.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Liu, H. X. and Feng, S. Curse of rarity for autonomous vehicles. *Nature Communications*, 15(1):1–5, 2024.

Liu, Z., Bahety, A., and Song, S. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023.

Lyu, Y., Luo, W., and Dolan, J. M. Risk-aware safe control for decentralized multi-agent systems via dynamic responsibility allocation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1–8, 2023.

Maggio, D. and Carlone, L. Bayesian fields: Task-driven open-set semantic gaussian splatting. *arXiv preprint arXiv:2503.05949*, 2025.

Maini, P., Goyal, S., Sam, D., Robey, A., Savani, Y., Jiang, Y., Zou, A., Fredrikson, M., Lipton, Z. C., and Kolter, J. Z. Safety pretraining: Toward the next generation of safe ai. *arXiv preprint arXiv:2504.16980*, 2025.

Margellos, K. and Lygeros, J. Hamilton–jacobi formulation for reach–avoid differential games. *IEEE Transactions on automatic control*, 56(8):1849–1861, 2011.

Mehta, S. A., Meng, F., Bajcsy, A., and Losey, D. P. Strol: Stabilized and robust online learning from humans. *IEEE Robotics and Automation Letters*, 9(3):2303–2310, 2024.

Nakamura, K., Peters, L., and Bajcsy, A. Generalizing safety beyond collision-avoidance via latent-space reachability analysis. *arXiv preprint arXiv:2502.00935*, 2025.

Nasiriany, S., Kirmani, S., Ding, T., Smith, L., Zhu, Y., Driess, D., Sadigh, D., and Xiao, T. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In *IEEE International Conference on Robotics and Automation*, pp. 8249–8257, 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Pandya, R., Liu, C., and Bajcsy, A. Robots that learn to safely influence via prediction-informed reach-avoid dynamic games. In *IEEE International Conference on Robotics and Automation*, pp. 14330–14337, 2025.

Peng, S., Chen, H., and Driggs-Campbell, K. Towards uncertainty unification: A case study for preference learning. *arXiv preprint arXiv:2503.19317*, 2025.

Ravichandran, Z., Robey, A., Kumar, V., Pappas, G. J., and Hassani, H. Safety guardrails for llm-enabled robots. *arXiv preprint arXiv:2503.07885*, 2025.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

Robey, A., Ravichandran, Z., Kumar, V., Hassani, H., and Pappas, G. J. Jailbreaking llm-controlled robots. In *IEEE International Conference on Robotics and Automation*, pp. 11948–11956, 2025.

Sagheb, S., Parekh, S., Pandya, R., Mun, Y.-J., Driggs-Campbell, K., Bajcsy, A., and Losey, D. P. A unified framework for robots that influence humans over long-term interaction. *arXiv preprint arXiv:2503.14633*, 2025.

Salzmann, T., Ivanovic, B., Chakravarty, P., and Pavone, M. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pp. 683–700, 2020.

Santos, L., Li, Z., Peters, L., Bansal, S., and Bajcsy, A. Updating robot safety representations online from natural language feedback. In *IEEE International Conference on Robotics and Automation*, pp. 7778–7785, 2025.

Seo, J., Nakamura, K., and Bajcsy, A. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. *arXiv preprint arXiv:2505.00779*, 2025.

Sermanet, P., Majumdar, A., Irpan, A., Kalashnikov, D., and Sindhwani, V. Generating robot constitutions & benchmarks for semantic safety. *arXiv preprint arXiv:2503.08663*, 2025.

Shi, Z., Zhao, E., Dennler, N., Wang, J., Xu, X., Shrestha, K., Fu, M., Seita, D., and Matarić, M. Hribench: Benchmarking vision-language models for real-time human perception in human-robot interaction. *arXiv preprint arXiv:2506.20566*, 2025.

Siciliano, B., Khatib, O., and Kröger, T. *Springer handbook of robotics*, volume 200. Springer, 2008.

Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., and Garg, A. Progprompt: program generation for situated robot task planning using large language models. *Autonomous Robots*, 47(8):999–1012, 2023.

Sinha, R., Elhafsi, A., Agia, C., Foutter, M., Schmerling, E., and Pavone, M. Real-time anomaly detection and reactive planning with large language models. *arXiv preprint arXiv:2407.08735*, 2024.

Song, C. H., Wu, J., Washington, C., Sadler, B. M., Chao, W.-L., and Su, Y. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2998–3009, 2023.

Sun, X., Meng, H., Chakraborty, S., Bedi, A. S., and Bera, A. Beyond text: Utilizing vocal cues to improve decision making in llms for robot navigation tasks. *arXiv preprint arXiv:2402.03494*, 2024a.

Sun, X., Zhang, Y., Tang, X., Bedi, A. S., and Bera, A. Trustnavgpt: Modeling uncertainty to improve trustworthiness of audio-guided llm-based robot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8794–8801, 2024b.

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.

Tian, L. and Oviatt, S. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction*, 10(2):1–32, 2021.

Tölle, M., Gruner, T., Palenicek, D., Schneider, T., Günster, J., Watson, J., Tateo, D., Liu, P., and Peters, J. Towards safe robot foundation models using inductive biases. *arXiv preprint arXiv:2505.10219*, 2025.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wabersich, K. P. and Zeilinger, M. N. Linear model predictive safety certification for learning-based control. In *IEEE Conference on Decision and Control*, pp. 7130–7135, 2018.

Wabersich, K. P., Taylor, A. J., Choi, J. J., Sreenath, K., Tomlin, C. J., Ames, A. D., and Zeilinger, M. N. Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, 43(5):137–177, 2023.

Wang, J., Sundarsingh, D. S., Deshmukh, J. V., and Kantaros, Y. Conformalnl2ltl: Translating natural language instructions into temporal logic formulas with conformal correctness guarantees. *arXiv preprint arXiv:2504.21022*, 2025a.

Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025b.

Wu, X., Chakraborty, S., Xian, R., Liang, J., Guan, T., Liu, F., Sadler, B. M., Manocha, D., and Bedi, A. S. On the vulnerability of llm/vlm-controlled robotics. *arXiv preprint arXiv:2402.10340*, 2024.

Wu, Y., Tian, R., Swamy, G., and Bajcsy, A. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. *arXiv preprint arXiv:2502.01828*, 2025.

Xing, S., Hua, H., Gao, X., Zhu, S., Li, R., Tian, K., Li, X., Huang, H., Yang, T., Wang, Z., et al. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *arXiv preprint arXiv:2412.15206*, 2024.

Xing, W., Li, M., Li, M., and Han, M. Towards robust and secure embodied ai: A survey on vulnerabilities and attacks. *arXiv preprint arXiv:2502.13175*, 2025.

Xu, C., Springenberg, J. T., Equi, M., Amin, A., Esmail, A., Levine, S., and Ke, L. Rl token: Bootstrapping online rl with vision-language-action models. *arXiv preprint arXiv:2604.23073*, 2026.

Xu, H., Koh, Y. S., Huang, S., Zhou, Z., Wang, D., Sakuma, J., and Zhang, J. Model-agnostic adversarial attack and defense for vision-language-action models. *arXiv preprint arXiv:2510.13237*, 2025.

Xu, R. and Ding, K. Large language models for anomaly and out-of-distribution detection: A survey. In *Findings of the Association for Computational Linguistics*, pp. 5992–6012, 2025.

Yang, Y., Duan, Z., Xie, T., Cao, F., Shen, P., Song, P., Jin, P., Sun, G., Xu, S., You, Y., et al. Fpc-vla: A vision-language-action framework with a supervisor for failure prediction and correction. *arXiv preprint arXiv:2509.04018*, 2025.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183, 2023.

A. Additional Deployment Examples for Modular Guardrails and Co-Design

This appendix provides two additional scenarios expanding on Sec. 6.3, illustrating how modular safety guardrails integrate with different FM-enabled robotics stacks and can be extended through cross-layer and cross-module co-design.

Example B: Open-Vocabulary Mobile Manipulation (Perception FM + Classical Planning/Control). A warehouse robot uses CLIP-based open-vocabulary detection (Gadre et al., 2023), along with a segmentation model e.g., SAM (Kirillov et al., 2023), for object identification, combined with classical motion planning and control. During a pick-and-place task, the perception FM may misidentify objects, for instance, confusing a fragile glass container with a plastic one, leading to inappropriate grasping force, or hallucinating object presence in cluttered scenes due to visual ambiguity. The Monitoring and Evaluation Layer generates trustworthiness scores by cross-validating detections against depth sensors and tracking consistency across frames; discrepancies (e.g., depth discontinuities inconsistent with detected object geometry) indicate potential misidentification. OOD detection methods (Farid et al., 2022) flag high-uncertainty cases such as novel object categories absent from training data or ambiguous boundaries caused by occlusion. Since no plans are produced by the FM-based component in this setting, the decision gate is optional and primarily used to prevent plans from relying on low-confidence perception outputs. The action gate provides execution-time enforcement (e.g., collision avoidance constraints), ensuring physical safety even when upstream perception is unreliable.

A co-design example of representation alignment: The key is to give the monitoring layer a representation that enables more informed downstream enforcement: instead of a scalar confidence, it outputs a pose-uncertainty ellipsoid that preserves the magnitude and direction of localization error. In low light, a low scalar score would otherwise force a blunt stop/go decision. With the ellipsoid, the action gate can enforce safety more precisely by tightening margins anisotropically (e.g., 8 cm along the major axis, 2 cm along the minor axis), and in future steps, the decision gate approves the same plan only if these ellipsoid-induced constraints remain feasible in the current workspace. This richer representation lets the guardrail be cautious where needed without resorting to uniform conservatism.

Example C: Dexterous Manipulation (End-to-end VLA Policy). A manipulation robot uses an RT-2-style VLA policy (Zitkovich et al., 2023) that maps visual observations and language instructions directly to control commands. This end-to-end design introduces distinct risks: perception or grounding errors can immediately translate into unsafe motion; the policy may hallucinate object locations under visual ambiguity or out-of-distribution scenes; and per-step "reasonable" commands can still accumulate into an unsafe path in clutter, gradually eroding clearance or steering the arm toward joint/workspace limits. Because the policy exposes few intermediate representations, failures are difficult to attribute (e.g., perception vs action prediction). The monitoring layer estimates action-level risk via token-level uncertainty from prediction entropy (Karli et al., 2025) and cross-validates scene understanding with an independent perception check against workspace constraints to flag hallucinations. Since no explicit plan is produced, the decision gate is bypassed and the action gate becomes the primary safeguard, projecting commands onto safe sets (Fisac et al., 2019; Ames et al., 2019) near collision or limit boundaries and triggering a fallback (e.g., controlled retraction) in high uncertainty. This setup mainly enforces action safety, with limited decision safety through uncertainty signals.

A co-design example of representation alignment and conservatism allocation: In the same VLA setting, the robot is told to "pick up the red cup" on a cluttered table near a glass vase; token-level entropy is high because the policy is unsure which candidate object to target. Without co-design, entropy is thresholded as a scalar, forcing a binary choice: execute with nominal constraints or fall back. With co-design, the monitor localizes uncertainty using attention/saliency and maps it into workspace coordinates as a spatial risk field, which the action gate uses to tighten margins and slow down only near high-risk regions. This turns uncertainty into localized, enforceable caution, enabling safe grasping without a hard stop.