Cross-Layer Adaptive Co-Design for Mission-Critical Embodied Intelligence

Rajat Bhattacharjya
Department of Computer Science; University of California, Irvine

PhD Forum Abstract

***Abstract*—Mission-critical embodied agents must sustain reliable perception and decision-making as missions, sensing conditions, connectivity, resources, and available human support change. Yet existing systems typically optimize isolated algorithms or execution choices, making otherwise capable pipelines brittle when these factors vary jointly. My research develops cross-layer adaptive co-design methods that coordinate algorithmic, architectural, communication, and runtime decisions according to mission and operating state. Across my completed works, adaptation moves progressively closer to the physical decision—from making computation feasible, to invoking it when valuable, matching execution to mission-required evidence, and validating whether prior actions remain safe to reuse—yielding up to 93.98% lower energy and 76.6% fewer safety violations. Building on this foundation, my dissertation advances a central premise: rather than directly selecting a model or system configuration, an agent should first express the next physical decision as an *evidence contract* specifying the information, fidelity, freshness, uncertainty, deadline, and risk it requires. A bounded metareasoning runtime will then compose computation, communication, validation, and assistance strategies to satisfy that contract according to their expected decision value, cost, and validity, and verify satisfaction before action. The intended outcome is a unified runtime that closes the loop from mission intent to sufficient evidence to trustworthy physical action.**

***Index Terms*—embodied intelligence, cross-layer co-design, adaptive runtime systems, metareasoning, human–robot collaboration**

I. Research Vision

Mission-critical embodied agents, such as disaster-response UAVs, exploration robots, and autonomous factory vehicles, must operate under tight energy, computation, and communication constraints. Their operating conditions are inherently variable: sensing may degrade, hazards may evolve, network support may disappear, and the role of a human operator may shift among supervision, intervention, and collaboration. These dependencies become increasingly consequential as embodied agents incorporate vision-language and vision-language-action (VLA) models. Representative VLA workloads target control rates of approximately 10–100 Hz, yet distributed inference latency can increase from 1.5 ms over 10-Gb Ethernet to 26 ms over 5G [1]. A model that is feasible under nominal conditions may therefore become too slow, energy-intensive, or unreliable when connectivity or resources change [2].

Resource variability is only part of the problem. The information required by an embodied agent also depends on the physical decision being supported. For example, a disaster-response operator may initially require rapid awareness of a developing hazard, but later require precise spatial evidence before sending a ground team through the affected region [3]. These mission needs impose different requirements on semantic content, spatial fidelity, freshness, confidence, latency, and risk. An efficient execution strategy is useful only if it produces the information needed for the current mission decision.

The systems problem is therefore not simply how to execute a model efficiently. An embodied agent must determine what evidence the next physical decision requires, which combination of sensing, computation, communication, and available support can produce that evidence, and whether the result is sufficient to justify action. Existing work addresses hardware-aware inference, distributed execution, clarification, and abstention [1], [4], but typically treats these as separate decisions. Open-world operation instead requires coordinating how evidence is produced, whether it is sufficient for action, and when external support is warranted.

My research addresses this challenge through *cross-layer adaptive co-design*: jointly adapting algorithmic, architectural, communication, and runtime decisions according to mission requirements, physical state, connectivity, and platform constraints. My work has demonstrated this progression in concrete systems, with the boundary of adaptation moving progressively closer to the physical decision—from determining which computations are feasible, to deciding when computation is valuable, matching execution to the information required by the mission, and verifying whether prior execution remains trustworthy.

Research hypothesis. I hypothesize that mission-critical embodied agents can achieve greater mission utility under fixed resource and risk budgets when a bounded metareasoning runtime identifies the evidence required to support the next physical decision and selects and composes cross-layer execution strategies according to the expected decision value of the evidence they produce, their execution cost, and their validity under the current operating state. Mission utility encompasses task success, timeliness, safety, and resource expenditure.

II. Cross-Layer Adaptive Co-Design

Fig. 1 summarizes how my work progressively expands the operating context considered by cross-layer adaptation. The common thread is a sequence of increasingly consequential execution decisions: which capabilities can execute on the available platform, when computation is valuable, what information the mission requires, and whether a previously successful decision remains valid. Each stage incorporates a richer view of the embodied system—from hardware constraints to physical dynamics, mission intent, connectivity, and execution state.

Co-designing resource-feasible intelligence. The first step is to avoid treating an algorithm as a fixed workload that must merely be mapped onto constrained hardware. `MUSIC-Lite` [5] exposes approximation opportunities in the singular-value-decomposition stage of MUSIC-based radar processing, reducing on-chip area by 17.25% and power by 19.4% with only 0.14% error. `HYPERDOA` [6] extends this principle beyond arithmetic approximation by replacing expensive matrix decompositions with hyperdimensional encoding and associative search. Under low-SNR coherent-source conditions, it achieves 35.39% higher accuracy than competing estimators while consuming 92.93% less energy than neural baselines.

The same principle extends from sensing algorithms to emerging AI workloads. `AAMLA` [7] predicts the GPU-memory requirements of alternative LLM adaptation methods, eliminates configurations that exceed the available budget, and selects among feasible methods before launching an autonomous hardware-generation workflow. `T-SAR` [8] complements this design-time capability through full-stack co-design of ternary representation, software dataflow, ISA support, and SIMD microarchitecture. Together, these systems expose which capabilities are feasible on constrained platforms and characterize their execution costs.

Adapting when computation is valuable. Making a capability feasible does not imply that it should execute continuously. `ACCESS-`

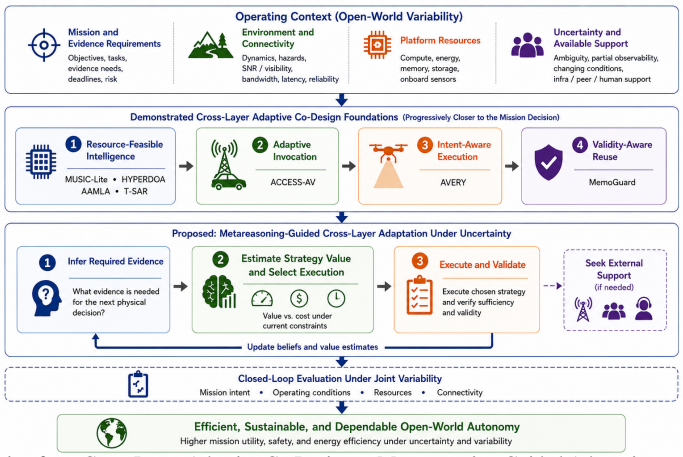


Fig. 1: Progression from Cross-Layer Adaptive Co-Design to Metareasoning-Guided Adaptation under Uncertainty.

AV [9] addresses this question for autonomous-vehicle localization by coupling the localization runtime to the physical system it observes. It opportunistically reuses periodically broadcast 5G synchronization signals and jointly adapts communication and computation according to signal quality and vehicle dynamics. Computation is therefore invoked according to the evolving need for a new position estimate rather than at a fixed design-time rate. This communication–computation–dynamics coupling maintains sub-30-cm localization accuracy while reducing average localization energy by 43.09%.

Matching execution to mission requirements. Physical dynamics can indicate when computation is valuable, but they do not determine what information the mission requires. AVERY [10] addresses this next question for distributed vision-language perception in disaster-response UAVs. It treats rapid contextual awareness and detailed spatial evidence as distinct semantic capabilities with different fidelity, freshness, computation, and communication requirements. The runtime uses operator intent to select the capability required by the current mission and then jointly adapts model partitioning, learned compression, and execution rate according to network conditions and onboard resources. AVERY achieves 11.2% higher accuracy than raw-image compression, and reduces onboard energy by 93.98% relative to full onboard execution.

Validating whether prior execution remains trustworthy. Selecting and efficiently executing an appropriate capability is insufficient when the assumptions supporting a prior decision no longer hold. MemoGuard [11] addresses this problem during communication loss, when a robot must determine whether previously successful experience remains safe to reuse. Retrieval similarity alone is insufficient: changed topology, battery margin, or outcome reliability can invalidate an otherwise relevant action.

MemoGuard inserts contract validation between episodic retrieval and action execution. A remembered action is reused only when its topology, resource, and outcome assumptions remain valid; otherwise, the system invokes local reasoning. It increases mission success from 32.6% to 84.2%, reduces safety violations by 76.6%, and avoids 21.4% of the fallback calls incurred by always reasoning.

Across these systems, adaptation successively determines what can execute, when execution is valuable, what information the mission requires, and whether prior execution remains trustworthy. Each decision, however, remains encoded in a specialized design or runtime. Open-world embodied intelligence requires a higher-level mechanism for coordinating these choices as mission requirements, uncertainty, resources, connectivity, and available support evolve together.

III. Conclusion and Future Work

Across my work, cross-layer adaptation has progressed from making computation feasible, to invoking it when valuable, matching execution to mission-required evidence, and validating whether prior decisions remain trustworthy.

Future work. The central premise of my dissertation is that an embodied agent should not adapt by directly selecting a model or system configuration. It should first express the next physical decision as an *evidence contract* specifying what information is required, at what fidelity, freshness, confidence, and risk. A contract-grounded, bounded metareasoning runtime will then treat computation, communication, prior experience, validation, and external assistance as composable strategies for satisfying that contract, selecting among them according to expected decision improvement, execution cost, and validity under the current state. After execution, the runtime will verify whether the contract has been satisfied and either permit physical action, acquire further evidence, or seek targeted agent, infrastructure, or human support. I will evaluate this approach through closed-loop disaster-response tasks that jointly vary mission requirements, sensing, connectivity, resources, deadlines, and assistance. The intended outcome is a unified runtime that closes the loop from mission intent to sufficient evidence to trustworthy physical action.

References

[1] W. Jiang *et al.*, "How fast can i run my vla? demystifying vla inference performance with vla-perf," *arXiv preprint arXiv:2602.18397*, 2026.

[2] R. Bhattacharjya *et al.*, "FRESHLATENT: Channel-aware latent adaptation for resource-constrained embodied VLM perception," 2026, arXiv:2609.30629. [Online]. Available: https://arxiv.org/abs/2609.30629

[3] R. Bhattacharjya *et al.*, "FloodReasonBench: Benchmarking VLM reasoning segmentation for embodied flood response at the edge," 2026, arXiv:2608.15410. [Online]. Available: https://arxiv.org/abs/2608.15410

[4] D. Yeke *et al.*, "The yes-man syndrome: Benchmarking abstention in embodied robotic agents," *arXiv preprint arXiv:2605.20544*, 2026.

[5] R. Bhattacharjya *et al.*, "MUSIC-Lite: Efficient MUSIC using approximate computing: An OFDM radar case study," *IEEE Embedded Systems Letters*, vol. 16, no. 4, pp. 329–332, 2024.

[6] R. Bhattacharjya *et al.*, "HYPERDOA: Robust and efficient DoA estimation using hyperdimensional computing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 20 841–20 845.

[7] R. Bhattacharjya *et al.*, "AAMLA: An autonomous agentic framework for memory-aware LLM-aided hardware generation," in *Proc. Int. Conf. VLSI Design (VLSID)*, 2026, pp. 171–176.

[8] H. Oh, K. Nam, R. Bhattacharjya *et al.*, "T-SAR: A full-stack co-design for CPU-only ternary LLM inference via in-place SIMD ALU reorganization," in *Proc. Design, Automation and Test in Europe (DATE)*, 2026, pp. 1–7.

[9] R. Bhattacharjya *et al.*, "ACCESS-AV: Adaptive communication-computation codesign for sustainable autonomous vehicle localization in smart factories," *ACM Transactions on Embedded Computing Systems*, vol. 25, no. 1, pp. 13:1–13:24, Jan. 2026. [Online]. Available: https://doi.org/10.1145/3771770

[10] R. Bhattacharjya *et al.*, "AVERY: Intent-driven adaptive VLM split computing via embodied self-awareness for efficient disaster response systems," 2025. [Online]. Available: https://arxiv.org/abs/2511.18151

[11] R. Bhattacharjya *et al.*, "MemoGuard: An adaptive runtime for guarding against memory traps in communication-limited robot navigation," 2026. [Online]. Available: https://arxiv.org/abs/2607.15589