Shi WS, Dong Z, Zhou PP. Physical Intelligence on the Edge: A vision for the decade ahead. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 2026. DOI: 10.1007/s11390-026-6292-8, CSTR: 32374.14.s11390-026-6292-8

Physical Intelligence on the Edge: A Vision for the Decade Ahead

Weisong Shi[1], *Fellow, IEEE, Distinguished Member, ACM*, Zheng Dong[2], and Peipei Zhou[3]

[1] *Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, U.S.A.*

[2] *Department of Computer Science, Wayne State University, Detroit, MI 48202, U.S.A.*

[3] *School of Engineering, Brown University, Providence, RI 02912, U.S.A.*

E-mail: weisong@udel.edu; dong@wayne.edu; peipei_zhou@brown.edu

Received January 10, 2026; accepted January 15, 2026.

Abstract This article examines key challenges in computing systems research under the emerging paradigm of Physical Intelligence on the Edge (PIE), in which raw sensor streams are transformed into real-time, safety-critical intelligence that can act in the physical world. It traces the evolution of computing architectures from centralized systems to distributed systems and edge computing, and argues that PIE constitutes a qualitative shift: the edge becomes the primary platform for tightly integrating sensing, reasoning, and actuation under stringent real-time constraints. The article identifies five emerging research thrusts—embodied spatial reasoning, embodied temporal reasoning, edge-native customization, symbiosis, and sustainability. Using a hypothetical PIE scenario, it exposes a fundamental gap between the capabilities of current systems and the requirements of future PIE-enabled autonomy: while today's edge platforms can execute individual components of perception and inference, they remain unable to autonomously close the sense-think-act loop with certifiable guarantees on timing and safety. This vision is further substantiated by recent industrial progress, including several compelling demonstrations showcased at CES 2026 by leading companies such as NVIDIA and AMD. The article concludes by calling for a paradigm shift in systems thinking—from efficiently transporting and processing data (bits) to predictably and safely influencing the physical world (atoms)—thereby positioning edge-native system design as a foundational enabler of next-generation autonomous and robotic systems.

Keywords embodied spatial/temporal reasoning, customization, symbiosis, sustainability

1 Introduction

Over the past decade, edge computing has emerged as one of the mainstream computing paradigms in modern computing systems, leading to the publication of several influential vision and survey papers[1–4] that have helped define the scope and foundations of the field. Among these, our visionary work[1] constitutes an early effort to articulate a coherent and unifying perspective. The central idea of this vision is a distributed computing paradigm that mitigates bandwidth and latency bottlenecks by placing computation in close physical proximity to data sources①. At that time, vision was largely defined by the movement of "bits": optimizing data flows between the cloud and the periphery. Although the architectural principles were sound, the realization of truly autonomous and intelligent action at the edge faced practical barriers; the hardware density and algorithmic efficiency required for general-purpose intelligence on constrained devices were not yet available. Unsurprisingly, early edge deployments emphasized caching, filtering, and narrow inference pipelines, falling short of autonomy[5].

In recent years of systems and hardware evolution, several capabilities that were largely considered exotic in 2016 have become commercially viable. These include heterogeneous accelerators[6–8] (e.g.,

Perspective

Special Issue: Celebrating the 40th Anniversary of JCST

① https://acm-ieee-sec.org/CSR-PI2018/report.html, Jan. 2026.

field-programmable gate arrays (FPGAs) and neural processing units (NPUs)), high-bandwidth, low-latency networks (e.g., 5G[9] and emerging 6G efforts), and large vision-language-action (VLA) models[10] capable of reasoning about physical context. Collectively, these advances position the community to pursue the next horizon of distributed systems: Physical Intelligence on the Edge (PIE). Formally, PIE represents an evolutionary stage of edge computing in which local, resource-constrained systems are endowed with the ability to perform autonomous, real-time, and safe physical actions by closing the sense-think-act loop in the physical world. Unlike traditional edge computing, which primarily optimizes data transport and preprocessing, PIE elevates actuation to a first-class systems objective, emphasizing bounded latency, predictability, and timely intervention in the environment. Fig.1 illustrates the historical citation trajectory of our visionary Edge Computing paper alongside its anticipated future impact, serving as a bellwether for an emerging research wave in which physical intelligence becomes a central organizing principle for next-generation edge systems.

Early embodiments of PIE are emerging across major sectors, including software-defined vehicles (SDVs)[11, 12], mobile manipulators in smart manufacturing[13], and multi-agent robotics in precision agriculture[14]. These domains differ in their sensing modalities, actuation dynamics, and safety envelopes, yet they share a common goal: closing the sense-think-act loop autonomously, reliably, and in real time, without constant reliance on the cloud.

The goal of this article is to clarify the computer-systems challenges posed by PIE. We begin by tracing the evolution from centralized computing to distributed systems and then to edge computing, and we argue why PIE introduces new requirements that are not reducible to traditional "move computation closer" narratives. Next, we sketch a hypothetical PIE scenario and ask why such applications remain largely experimental rather than ubiquitous today. From that starting point, we examine key research problems, including edge-native hardware-algorithm co-design[15, 16], certifiable real-time autonomy[17], and robust operation under open-world uncertainty[18]. To preserve focus, we avoid digressions into algorithmic advances in AI per se, and instead emphasize operating systems, architecture, and hardware abstractions necessary to make physical intelligence a dependable systems reality.

2 Evolution

PIE represents a major step in a lineage that extends from centralized mainframe computing[19] through distributed systems[20] and edge computing[1]. Some technical issues in PIE correspond to problems already studied earlier in this evolution; in some cases, existing solutions apply directly. In other cases, the demands of the physical world, including irreversibility, safety envelopes, and strict timing, alter the problem sufficiently that new solutions are required. PIE also introduces problems that do not map cleanly onto prior systems models. In the rest of this section, we develop a taxonomy of issues characterizing each phase of the evolution, highlighting both

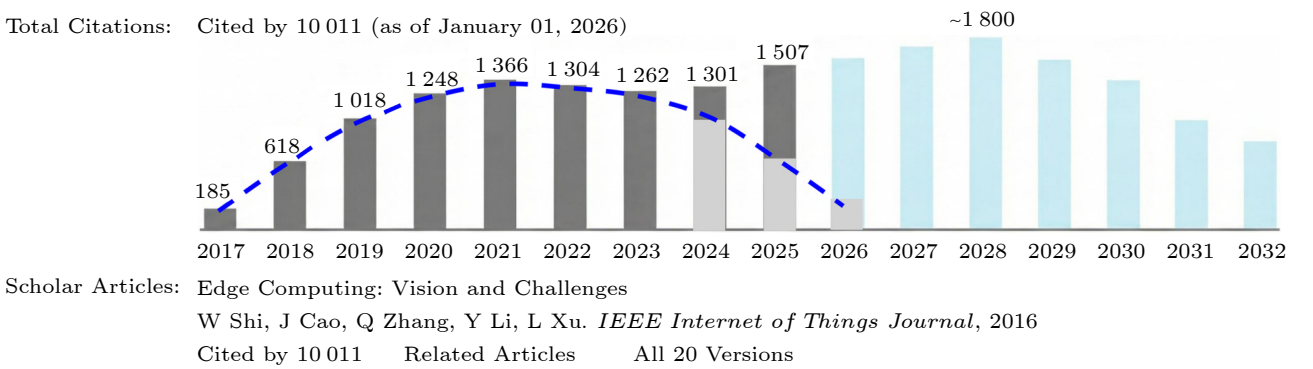


Fig.1. Citation trajectory of our 2016 visionary paper on edge computing[1]. Annual citation counts from 2017 through 2025 capture the first decade of its citation history, as reported by Google Scholar. While absolute citation counts may vary across databases (e.g., Web of Science or Scopus), the overall trends remain consistent. In many research domains, citation dynamics exhibit an approximately ten-year lifecycle, evolving from initial emergence to maturity and eventually tapering off, as suggested by the dashed trend line. Edge computing, however, deviates markedly from this canonical pattern. Rather than plateauing, citation activity has continued to accelerate, with particularly strong growth observed in 2024 and 2025 and a projected peak around 2028. We therefore anticipate a renewed surge in citations over the next three to four years, driven by the rise of Physical Intelligence on the Edge. The projected citation counts are indicated by the light-blue bars.

continuity and discontinuity.

2.1 Centralized and Distributed Systems

The field of computer systems initially developed under a centralized model, in which information was stored and computed exclusively on a server while clients acted as passive terminals. Research in this era centered on efficient retrieval, access control, and the consistency of centralized data stores. As data volume and user demand increased, however, the single-server model became untenable. This scalability pressure catalyzed the emergence of distributed systems, whose algorithmic foundations, including remote communication, fault tolerance, replication, and high availability, remain essential today[21]. Yet even in distributed systems, the dominant objective remains information intelligence: managing and moving bits between storage and users, with correctness defined primarily over digital state.

2.2 Edge Computing

The explosion of data volume and the proliferation of high-rate sensors in the early 21st century exposed limitations of cloud-centric information intelligence: in many applications, data could not be transported and processed fast enough when the compute substrate was geographically distant[22–24]. Edge computing thus emerged to move computation closer to data generation. While many distributed-systems principles continued to apply, the primacy of latency motivated specialized techniques such as computation offloading, localized caching, and bandwidth-adaptive processing. In many deployments, however, edge computing still treated data as something to be filtered, compressed, or sparsely inferred upon, prioritizing responsiveness rather than supporting safe, autonomous physical intervention.

2.3 PIE

PIE goes beyond edge computing by elevating physical action, rather than low-latency inference alone, to a first-class systems outcome. As modern AI models improve, an edge device can increasingly progress from processing signals to extracting semantics and making decisions. Yet PIE requires more than "smarter inference at the edge": it requires systems that can close the sense-think-act loop under explicit constraints imposed by physics, safety, and energy.

Accordingly, PIE expands the edge-computing agenda along the following five research thrusts.

Embodied Spatial Reasoning. A physical environment is not merely a stream of measurements, but a dynamic three-dimensional (3D) reality governed by geometry and physics[25]. Embodied spatial reasoning integrates perception with physical control, enabling the system to reason not only about "what" an object is, but "where" it is relative to the body, how it can be contacted, and how it will behave under manipulation. For example, a mobile manipulator may adapt grasp force based on inferred friction and compliance, while contact feedback can immediately refine the internal map when vision is ambiguous. Complete physical omniscience is unrealistic, but functional spatial competence, achieved through the tight coupling of vision, proprioception, and tactile sensing, is well within reach.

Embodied Temporal Reasoning. In the physical world, time is not merely a performance metric; it is a boundary condition for safety and stability[26]. Cloud inference pipelines often optimize for average-case throughput and tolerate long-tail latency. In PIE, the latency of the perception-to-actuation path must be bounded; otherwise, the underlying dynamics can destabilize the system, with potentially catastrophic outcomes. This elevates worst-case reasoning (e.g., bounding end-to-end delay and jitter) to a central design goal. While strict determinism may be unattainable for complex stochastic models, certifiable timing bounds and safe fallback behavior are plausible targets for systems design.

Edge-Native Customization. PIE workloads vary dramatically across environments and missions. Edge-native customization tailors the hardware-software stack to the operating context, bridging the gap between general capability and application-specific utility[27]. For instance, an autonomous container carrier in a controlled smart port may prioritize energy efficiency and long endurance, whereas a rescue vehicle in post-earthquake debris may prioritize robust perception and aggressive uncertainty handling, even at high energy cost. A key research direction is dynamic reconfiguration: the ability to adjust sensing fidelity, model complexity, and actuation policies as mission context changes.

Symbiosis. PIE is not fundamentally about replacing humans; it is about designing collaborative sys-

tems in which biological and silicon intelligence complement one another. In practice, PIE systems may automate high-volume repetitive physical tasks with precision, while humans provide judgment under rare, high-stakes edge cases[28]. Such division of labor resembles aviation: autopilot manages routine control, while pilots intervene under severe turbulence or anomalous conditions. Designing this partnership requires systems support for transparent intent, controllable autonomy, and principled hand-offs.

Sustainability. PIE will be deployed under widely varying energy constraints and carbon budgets[29]. Grid-connected industrial arms and battery-powered drones inhabit different feasibility regimes, and energy scarcity can be fatal to long-duration autonomy. Sustainability therefore becomes a systems objective: dynamically trading off intelligence, sensing fidelity, and actuation aggressiveness against energy availability. For example, a mobile agent might switch to a smaller perception model or reduced sensing rate when battery is critical to ensure safe return-to-base. Complete energy independence is unlikely; sustainable operation through intelligent resource management is a practical and pressing goal.

2.4 PIE vs Embodied AI

While PIE and Embodied AI[30] share the ultimate vision of agents that interact intelligently with the physical world, they address different layers of the realization stack. Embodied AI primarily focuses on the "algorithmic" capability, asking "how can an agent learn to perceive and act?", often abstracting away the underlying computational costs. In contrast, PIE focuses on the "systems" infrastructure required to sustain that behavior in the real world, asking "how can we execute this action safely within power and latency budgets?".

This distinction manifests in three key dimensions.

• *Algorithms vs Infrastructure.* Embodied AI research typically prioritizes the learning of robust policies and representations, often utilizing simulation where resources are abundant. PIE treats the AI model as a workload to be managed, focusing on the hardware-software co-design required to deploy these models on constrained edge devices. Where Embodied AI aims for high task success rates, PIE aims for operational feasibility, ensuring that the heavy computational demands of the AI do not exceed the thermal, energy, or form factor limits of the physical host.

• *Logical Time vs Physical Time.* In many Embodied AI paradigms, particularly reinforcement learning simulations, time is treated as a discrete logical sequence $(t \to t+1)$; the environment waits for the agent to compute. In PIE, time is a strict, continuous boundary condition. The system must guarantee that the total latency of the sense-think-act loop is less than the stability margins of the physical dynamics. Consequently, PIE prioritizes worst-case execution time (WCET) and real-time scheduling rather than the average-case throughput commonly accepted in standard AI inference.

• *Static vs Dynamic Constraints.* Embodied AI models often assume a fixed computational budget. PIE, driven by the Sustainability and Edge-Native Customization thrusts, views resources as dynamic. A PIE system must actively trade off algorithmic fidelity against survival, potentially reverting to simpler perception models or lower control frequencies to extend battery life or reduce heat, a form of systems-level optimization rarely addressed in pure Embodied AI research.

3 An Example Scenario

What would it mean to operate within a world enabled by PIE? To characterize the resulting system behavior and user experience, we construct a hypothetical but technically plausible scenario set in 2035. Although the scenario instantiates PIE as the underlying systems paradigm, the architectural principles and mechanisms it highlights are broadly applicable across emerging computing systems.

Alice stands at the entrance of a bustling shopping mall in Osaka, Japan, searching for premium sashimi. It is late afternoon, and she does not speak the local language. She signals a passing autonomous service cart, a specialized unit designed to maneuver through the mall's fixed topology while remaining robust to the dynamic flow of guests. Her smartphone transfers her semantic intent and dietary profile, captured via a wearable device, to the cart, allowing it to act as her agent. By combining the mall's live inventory feeds with Alice's personalized preference history, the system infers an appropriate vendor and guides her through crowded corridors, pre-translating menu information onto her AR (augmented reality) glasses.

Alice arrives just as the day's fresh catch is placed on display. As the transaction unfolds, she reaches for a pre-packaged assortment containing shellfish. Her AR glasses detect a hazard: Alice's records indicate a

severe shellfish allergy. The system issues a haptic alert and overlays a conspicuous warning, and Alice withdraws her hand and selects tuna instead. As she exits, the mall begins to close. Alice watches the cart autonomously navigate to a docking station, where it swaps its interactive interface for a sanitation module and begins overnight cleaning. The same embodied platform thus serves as a daytime guidance agent and a nighttime maintenance agent, improving both safety and utilization.

4 Gaps in the State of the Art

This scenario illustrates several core ideas in PIE. It demonstrates "symbiosis": Alice can navigate a foreign environment and transact safely because the system acts as her semantic agent and intervenes when risk is detected. It also demonstrates "edge-native customization": the cart is not a generic rover, but a design optimized for a particular facility while remaining resilient to crowd dynamics. Finally, it demonstrates "sustainability" through dual use: the platform shifts from daytime assistance to nighttime sanitation, maximizing the utility extracted from embodied hardware and its associated energy and carbon costs.

The scenario also highlights cross-layer integration. Inventory feeds are infrastructure-level signals; dietary constraints are user-level semantics; "after-hours" is facility-level context. Only by composing these disparate sources into a coherent state can a PIE system both assist the user and maintain the environment.

Perhaps the most striking aspect of the scenario is that many component technologies exist in isolation today. The hardware (e.g., autonomous carts, sanitizing attachments, AR glasses, haptic interfaces) is commercially plausible, and the software components (e.g., translation, intent inference, SLAM (simultaneous localization and mapping)) have been demonstrated. Why, then, does the end-to-end experience still feel like science fiction? The answer is that the whole is greater than the sum of its parts. The primary gap is not a missing model or a missing sensor, but the absence of a unified, edge-native architecture that can reliably close the sense-think-act loop with strong timing predictability and certifiable safety. Current systems can run the cart or run the translation, but they rarely provide principled guarantees that the integrated loop will remain safe, timely, and robust across changing conditions, without constant human supervision.

5 Charting the Road Ahead

Realizing PIE in practice requires addressing a range of difficult design and implementation problems. Building on the discussion above, we now examine a set of architectural challenges at finer granularity. Our aim is not to be exhaustive, but to convey a representative view of the road ahead. The topics discussed here are therefore a selective sampling of the broader problem space, with no intended ordering or claim of exclusiveness.

We assume that each user is surrounded by a continuous, agentic computing sphere that accompanies them and mediates interactions with nearby digital and physical infrastructure. Importantly, this mediation extends beyond passive information processing: the system can initiate and regulate physical actuation in the world. This personal sphere is likely to emerge as a distributed constellation of heterogeneous devices, ranging from body-worn biosensors to augmented-reality eyewear, that collectively operate as a single coherent entity.

We refer to this entity as a ''physical agent'' of the user, deliberately distinguishing it from the passive ''client'' model that characterizes traditional edge computing. Unlike a client that retrieves or caches data, a PIE agent reasons about physical causality: geometry, force, risk, and timing. It can coordinate and safely manipulate external embodied platforms (e.g., an autonomous service cart) or instrumented environments (e.g., smart doors and elevators). Supporting these capabilities requires substantial systems sophistication and, consequently, increased complexity.

Fig.2 illustrates the systematic structure of a representative PIE agent as a concrete example of this complexity. Beyond conventional modules for sensing, communication, and data processing, the architecture incorporates components required for moving and agentic operations. While legacy modules for wireless data transmission remain essential, new components, such as Environmental Understanding and Hazard Detection, are introduced to support safe physical-world operations. As requirements for predictable physical intervention become better understood, additional components will likely emerge.

5.1 AI Adoption for PIE

Adopting AI in PIE is not primarily a question of whether modern foundation models are "capable"; it is a question of whether their capabilities can be oper-

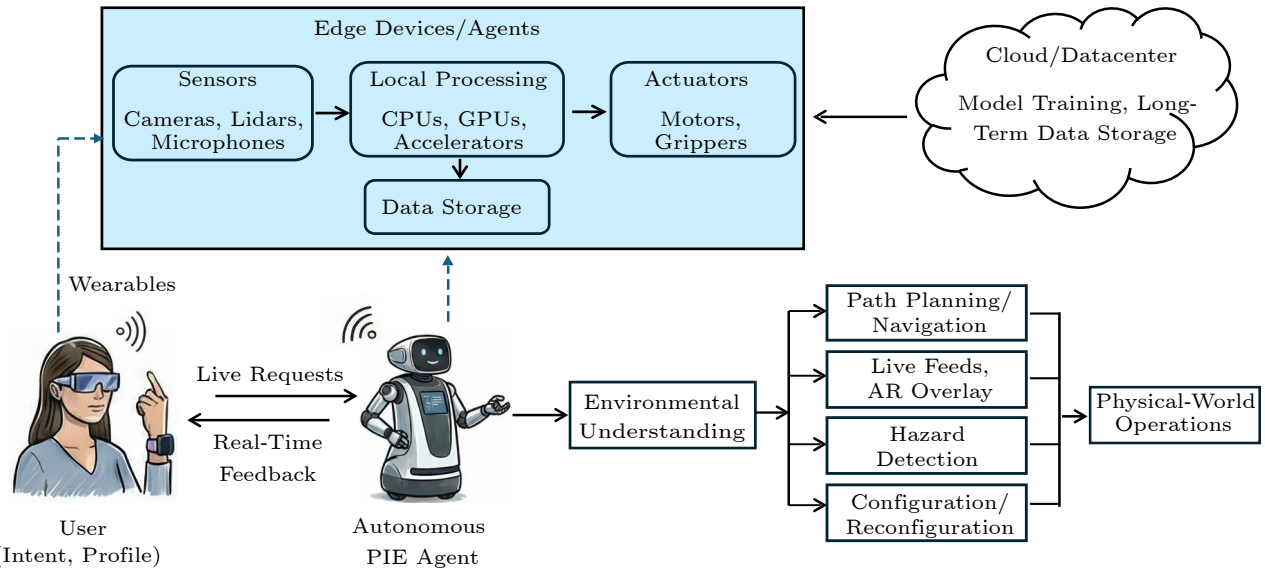


Fig.2. Key components of PIE agents running on edge devices that enable physical edge intelligence for real-world interaction and autonomous agentic operations, together with the critical software functions executed by the PIE agent in response to user requests.

ationalized on resource-constrained edge platforms that must act under bounded latency[31], strict energy budgets[32], and certifiable safety envelopes. Today's AI state of the art is increasingly concentrated in data centers, where large attention-based models achieve impressive semantic competence. PIE, however, requires these models to participate in a closed sense-think-act loop, where a missed deadline is a timing fault and a wrong action can be physically irreversible. This mismatch creates a systems gap: current AI is powerful but difficult to make predictable, auditable, and sustainable at the functional edge, such as robots, vehicles, and wearable proxies.

For example, consider an autonomous service cart that must 1) interpret a user's intent, 2) plan a safe path through a dense crowd, and 3) intervene in real time if a hazard is detected, such as an allergy risk. A cloud-scale model can often reason about the situation in language, but it may not respond within a bounded deadline, and it may propose an action that is socially plausible yet physically unsafe, such as politely yielding to a pedestrian gesture even when braking distance is insufficient for an autonomous vehicle, or attempting to move through a dense crowd in a courteous manner without accounting for sensor latency and minimum separation constraints in an autonomous service cart. Conversely, a small edge model can meet latency targets, but may fail under open-world novelty. Bridging this gap requires an architectural approach to AI adoption rather than a single "best model" deployment.

5.1.1 *Limitations of Current Systems*

Today's AI deployments at the edge typically fall into one of two unsatisfying extremes.

On the one end are narrow edge pipelines, consisting of lightweight perception or detection models coupled to fixed heuristics[22, 33]. These systems can be fast and power-efficient, but their intelligence is brittle. They fail when the environment shifts, when novel objects appear, or when intent requires multi-step reasoning. Such pipelines can recognize and react, yet they rarely compose capabilities into robust autonomy.

At the other extreme are monolithic foundation-model stacks that provide impressive open-world reasoning, but whose resource profile and timing behavior are poorly matched to PIE. Attention-based models remain expensive in memory bandwidth and compute, and their end-to-end response time is vulnerable to long-tail latency and thermal throttling[34, 35]. More critically, their outputs are difficult to certify, since the reasoning chain can be opaque, input-dependent, and not easily auditable against safety constraints. In embodied settings, this is not merely inconvenient, but it can also be unsafe.

Complicating matters further, current AI tooling assumes that "intelligence" is executed as an isolated inference call, whereas PIE requires continuous operation, with high rate sensor streams, tight actuation deadlines, and correctness defined over closed-loop behavior rather than static predictions. This exposes a fundamental architectural deficit: we lack edge-native abstractions that make learning-based intelligence sche-

dulable, compositional, and monitorable in real time.

5.1.2 Open Research Questions

• *Composable Agentization.* How should intelligence be divided into specialized agents built on small reasoning foundation models? Which functions should reside at the edge, such as reflex safety, local navigation, and intent parsing, and which should be delegated to a large foundation model controller via explicit escalation behavior?

• *Tool-First Execution.* How can agents be constrained to act through structured, typed function calls, rather than free-form generation? What is the right interface boundary where the model proposes, but deterministic code executes, validates, and logs?

• *Semantic Agent Communication and Explainability Verification.* Can we define a probabilistic process-calculus-based protocol for agent-to-agent communication that carries typed intents, uncertainty, and explicit preconditions and postconditions? How can such messages be used to support explainability verification, that is, auditing that an action follows from admissible semantic commitments?

• *Data Production and Consumption Policies.* How should PIE systems govern data flows when consumers must sometimes prioritize policy consistency, such as safety, privacy, authentication, and at other times prioritize availability, such as continued operation under disconnection? Can federated data models with data virtualization and common semantic classifiers prevent brittle coupling across heterogeneous sensors and infrastructure feeds?

• *Resource Consumption Beyond Throughput.* Since resource consumption is still a binding constraint for physical AI, how should the stack minimize reliance on attention transformers in the real-time critical path? What near-term synthesis techniques can produce more specialized yet more efficient edge agents, and what future model families, such as state-space or flow-based models, could provide better latency and energy scaling?

• *Causal-Conceptual Safety Interfaces.* Can we construct a formally checkable safety layer that maps input tokens/signals to concept triggers, composes these concepts semantically, and screens candidate actions using logical calculus to predict, or bound, behavioral outcomes? Which properties can realistically be verified at runtime, and which must be verified offline?

• *Learning Without Heavy Supervision.* Since supervised learning is often effective only for low-level mappings, how should PIE systems integrate reinforcement learning (RL)[36] and semi-supervised or self-supervised methods while preserving safety and predictability? What systems mechanisms are required for safe policy updates, including versioning, rollback, runtime gating, and simulation-to-real auditing?

5.2 Semantic-to-Physical Translation

For PIE to be effective, the system must decouple a user's high-level semantic goal from the brittle specifics of physical execution. Otherwise, the system becomes fragile, failing whenever ideal physical conditions are not met.

For example, suppose a user expresses a desire to drink water. The system initially plans to fetch a glass. However, sensors detect that all glasses are currently in the dishwasher. At this point, should the system:

• terminate the task and report "Object Not Found"?

• suspend the task indefinitely until a glass becomes available?

• reason about affordances, identify a clean ceramic bowl as a viable substitute for holding liquid, and deliver it to the user?

The correct response depends on whether the system recognizes that the underlying goal is hydration, not the acquisition of a specific geometric cylinder.

5.2.1 Limitations of Current Systems

Today's systems struggle to separate semantic goals from concrete object instances[37, 38]. On one end are rigid command-and-control pipelines that map a request such as "get water" directly to a fixed action, e.g., `find(cup)`. When the cup is unavailable, the reasoning process collapses. On the other end are purely generative models, which may propose a bowl as a substitute at the language level, yet lack embodied mechanisms to verify whether a particular bowl is clean, graspable, stable, and socially appropriate for the user. Bridging this gap by grounding abstract goals in a flexible physical reality raises foundational systems questions.

5.2.2 Open Research Questions

• *Intent Hierarchies.* Can the system distinguish

between a goal (e.g., drinking water) and a method (e.g., using a cup)? Does it maintain a fallback ontology that links the goal of drinking to any object capable of liquid containment?

• *Affordance-Based Representations.* How are objects represented internally? Are they identified solely by semantic labels (e.g., "cup", "bowl"), or by affordances such as containment, graspability, thermal insulation, and cleanliness? Can the system infer that a bowl supports hydration even if it has never observed a human drinking from a bowl?

• *Norms, Preferences, and Confirmation.* How should the system handle deviations from social norms or user expectations? While a bowl satisfies the physical requirement, it may violate etiquette or preference. At what level of deviation (e.g., measuring cup versus flower vase) must the system request confirmation?

• *Timing of Substitution.* Does searching for substitutes introduce unacceptable delay? Can the system scan the environment, identify viable alternatives, and synthesize a new grasp-and-delivery plan quickly enough to satisfy an immediate need under real-time constraints?

5.3 Navigating the Physical World

5.3.1 Challenges

Navigation in the physical world poses fundamental challenges for PIE agents[39–41]. Unlike purely computational environments, physical navigation requires action under incomplete information, strict timing constraints, and irreversible consequences. In this setting, the limiting resource is often not peak compute, but the fidelity and timeliness of the agent's internal world model relative to environmental dynamics.

First, navigation is constrained by imperfect and heterogeneous sensing[42]. Vision collapses a 3D world into two-dimensional (2D) projections, often producing "too much information" without sufficient depth or physical semantics. Navigation-relevant properties such as surface geometry, compliance, friction, and load-bearing capacity are difficult to infer from vision alone. Tactile and proprioceptive sensing therefore become essential, yet they introduce the "hands problem", where certain information can only be obtained through physical contact. Safe navigation thus requires tight sensor fusion, including reliable hand-eye coordination that aligns visual observations with contact-based feedback.

Second, navigation is intrinsically distributed and time-sensitive[43]. Mobile agents must move, localize, and coordinate while relying on wireless communication with latency, jitter, and intermittent connectivity. PIE nodes must therefore maintain correct action sequencing under non-ideal communication. Classical distributed systems issues reappear in physical form: non-deterministic state evolution, idempotent coordination, and robustness to delayed or duplicated messages. In navigation, however, a delayed or repeated "message" may translate into mistimed motion, directly impacting safety.

Third, navigation must explicitly account for uncertainty and irreversibility[44]. Real environments are partially observable and change unpredictably, making purely deterministic planning insufficient. Navigation therefore demands hierarchical planning and bounded probabilistic reasoning with explicit safety margins. Unlike software operations, physical actions cannot always be undone: collisions, falls, and damage are irreversible outcomes. A navigating agent must reason not only about optimality, but also about risk, deciding when to proceed cautiously, execute a fail-safe maneuver, or halt.

These challenges underscore why navigation intelligence in PIE must be edge-centric. Continuous sensing, mapping, and motion planning generate high-rate streams that cannot be fully offloaded to the cloud without violating latency and safety constraints. At the same time, the scarcity of task and environment-specific training data in open-world settings limits the effectiveness of purely data-driven approaches. Robust navigation thus demands co-designed sensing, computation, and learning mechanisms that operate locally.

5.3.2 Open Research Questions

• How should a PIE agent fuse visual, tactile, and proprioceptive signals, such as joint positions, velocities, and force or torque feedback that reflect the system's internal state, to maintain a reliable world model in real time?

• How should conflicting sensory cues be resolved during motion?

• What are the costs of constructing and updating 3D semantic maps fast enough to support safe navigation?

• Furthermore, is physical navigation primarily a networking problem (coordination over wireless links)

or a systems problem in which timing guarantees are inseparable from safety?

• How should idempotence be defined when repeating an action may cause a collision or fall?

• Finally, what probabilistic bounds are acceptable when humans are nearby? Should an agent freeze under extreme uncertainty, or attempt a constrained fail-safe motion?

Addressing these questions is central to enabling reliable navigation in the physical world and, more broadly, to realizing the vision of PIE.

5.4 Timing Predictability

5.4.1 Limitations of Current Architectures

Temporal predictability becomes essential when there is a mismatch between computational execution time and the rigid dynamics of the physical world. In such systems, the critical resource is not average throughput but worst-case latency and jitter[45]. The consequences of timing unpredictability differ fundamentally depending on whether intelligence resides in the cloud or at the edge.

First, consider traditional cloud-based inference, where massive computational resources are available but hard temporal guarantees are absent. This approach suffers from the well-known long-tail latency problem[46]. While 99% of requests may return within 50 ms, the remaining 1% can experience delays of hundreds of milliseconds due to network congestion, packet loss, or server-side queuing. In purely digital settings (e.g., loading a web page), such delays are inconvenient; in PIE, they can be catastrophic. A robot balancing on two wheels cannot wait 500 ms for a stabilization update without risking a fall.

Second, PIE pushes time-critical actuation loops to the local edge. By removing the wide-area network from the critical path, variance in message delivery is reduced, enabling bounded response times (deadlines) for actuation. Unlike cloud systems that optimize for throughput (tasks per second), PIE systems must optimize for jitter minimization and deadline adherence, emphasizing consistency in completion times[47].

Third, hierarchical architectures can combine these regimes[48]. The edge executes high-frequency reflex-like safety loops under hard real-time constraints, while the cloud performs low-frequency long-horizon planning and learning under soft real-time constraints. This organization mirrors biological systems: the spinal cord manages immediate reflexes (fast and predictable), whereas the brain handles complex reasoning (slower and more variable).

Collectively, these strategies underscore the central role of timing in physical intelligence. PIE depends on hardware accelerators and real-time operating systems capable of executing perception and decision models within strict temporal bounds. While the cloud remains valuable for non-critical learning and adaptation, immediate control of physical systems requires certifiable predictability and principled fallback behavior.

5.4.2 Open Research Questions

Despite recent progress, fundamental questions remain.

• How can one rigorously bound the Worst-Case Execution Time (WCET) of deep neural networks whose execution may be data-dependent?

• Can we construct safety envelopes that guarantee a decision within N milliseconds regardless of input complexity, possibly via degraded-but-safe modes?

• While moving computation to the edge reduces network-induced variance, is this still true under tight energy constraints? Does thermal throttling introduce new forms of temporal unpredictability?

• How should a PIE system handle timing faults? If a deadline is missed due to transient overload, should the system attempt recovery, trigger an immediate mechanical fail-safe, or switch to a certified fallback controller?

• Is the "simulacrum" of real-time offered by 5G/6G sufficient for safety-critical physical AI? Can mechanisms such as network slicing ensure deterministic delivery for reflex loops, or must time-critical intelligence remain physically on the machine?

5.5 Energy Efficiency

Energy efficiency is a first-class systems constraint that shapes what intelligence can be executed, how long autonomy can be sustained, and even whether timing guarantees remain valid. In contrast to cloud-scale AI, where power delivery and cooling can be provisioned as infrastructure, PIE systems operate under tight and often non-negotiable energy envelopes: battery-powered wearables, service carts with limited duty cycles, drones with minutes of flight time, and mobile manipulators that must share power budget across computation, sensing, communica-

tion, and actuation. In such settings, the marginal cost of intelligence is not abstract compute, but reduced endurance, thermal throttling, and degraded safety margin.

Two properties make energy particularly challenging for PIE. First, energy consumption is coupled to timing predictability. As power draw rises, thermal constraints trigger dynamic voltage frequency scaling (DVFS) and throttling[49], thereby increasing latency and jitter when bounded response time is essential. Second, energy is coupled to physical action. A system that conserves compute energy but wastes actuation energy, e.g., through inefficient motion planning, excessive braking or acceleration, or repeated retries due to perception failures, may still be unsustainable. Thus, PIE requires a holistic view of energy that spans compute, sensing, networking, and mechanics.

5.5.1 *Limitations of Current Systems*

Current edge stacks largely treat energy as an afterthought or as a single-layer optimization problem. Many deployments depend on model compression, including quantization and pruning, together with hardware accelerators[50], yet these techniques do not guarantee good system-level energy behavior. For instance, a compressed model may reduce MAC operations but introduce irregular, non-consecutive memory accesses, yielding limited improvement on real devices. Similarly, executing a large model intermittently may appear efficient, but bursty inference can induce thermal spikes that violate real-time guarantees during subsequent control cycles.

Complicating matters further, energy management today is often decoupled from mission intent. A PIE agent might lower frame rate to save energy without realizing that the environment has become crowded and risk has increased. Conversely, it might maintain a high-fidelity perception pipeline even when the task is low stakes, e.g., escorting a user through an empty corridor. Without intent-aware control of energy, systems either waste energy or sacrifice safety.

5.5.2 *Open Research Questions*

- *Energy-Timing Co-Guarantees.* How can one provide joint guarantees of bounded latency and bounded energy under thermal constraints? Can schedulers incorporate energy as a first-class resource alongside time and bandwidth, producing energy-aware deadlines that trigger safe degradation before thermal throttling occurs?
- *Cross-Modal Energy Allocation.* How should a PIE node allocate energy across sensing (e.g., camera/LiDAR/tactile), compute (e.g., NPU/GPU/CPU), communication (e.g., 5G/Wi-Fi), and actuation? Can the system maintain an explicit value-of-information model that determines when additional sensing is worth its energy cost?
- *Energy-Proportional Intelligence.* Can we build multi-resolution, multi-model stacks where intelligence scales smoothly with energy budget, e.g., switching between small edge agents and larger reasoning modules without destabilizing closed-loop control? What is the right granularity of switching to avoid oscillation?
- *Embodied Energy Accounting.* How should energy accounting include the cost of physical motion, retries, and safety maneuvers? Can planners jointly optimize for risk and energy, producing trajectories that are not only safe but also energy-stable over long operation?
- *Sustainable Lifecycle Operation.* Beyond runtime energy, PIE sustainability also includes embodied carbon and device lifetime. How should the stack incorporate hardware wear, battery aging, and maintenance scheduling, particularly in fleet settings like mall carts that must operate continuously with predictable availability?

5.6 Privacy and Trust

PIE systems exist at the boundary between private human life and public physical space. A PIE agent must observe the world to act safely, but observation itself can be invasive: continuous camera streams, biometric signals from wearables, identity-linked intent histories, and location traces inside shared environments. Unlike conventional edge applications that process data for convenience, PIE processes data to intervene in the physical world, making privacy failures and trust breakdowns not only informational harms but also potential safety hazards. If a user does not trust the system, they will disable it; if the system cannot authenticate its partners, it will act on adversarial signals; if the environment cannot verify the system's authority, it will reject legitimate actions. Trust should therefore be viewed not as a social afterthought, but as a fundamental systems dependency.

In PIE, privacy and trust also interact with timing. Cryptographic verification, secure logging, and policy enforcement introduce overhead. If security is bolted on naively, it can increase latency and jitter in the sense-think-act loop. Conversely, if security is weakened to meet deadlines, the system becomes vulnerable precisely because it is time-critical. The core challenge is therefore to design real-time compatible privacy and trust mechanisms.

5.6.1 Limitations of Current Systems

Today's privacy protections are largely designed for either 1) cloud-centric services where computation is remote and policy enforcement is centralized, or 2) mobile-device settings where actions are mostly digital. PIE breaks both assumptions. Data is multi-party (e.g., user, facility, vendors), multi-modal (e.g., vision, biosignals, infrastructure feeds), and continuous, and actions are physical. Existing permission models struggle to express contextual policies such as: "allow allergy detection in the mall, but do not store raw video", or "permit the cart to receive my dietary constraints only for this transaction window".

Furthermore, current defensive postures are largely reactive and passive. When a PIE agent, such as an autonomous vehicle, detects a physical-world attack or an anomalous environment, the default "safe" behavior is typically to halt or stay in place. While this minimizes immediate kinetic risk, it leaves the agent vulnerable to persistent threats or entrapment. Current systems lack the adversarial awareness required to distinguish between a mechanical failure and a targeted physical intervention, limiting their ability to execute evasive or protective maneuvers.

5.6.2 Open Research Questions

- *Policy-Grounded Data Minimization.* What should be the default data representation in PIE: raw sensor streams, features, or semantic commitments? Can we build a pipeline that transforms raw observations into minimal sufficient representations for action, and provably discards what is unnecessary?
- *Real-Time Secure Execution.* How can trusted execution environments, secure boot, and attestation be integrated without violating timing guarantees? Is there a principled separation between hard real-time safety loops and soft real-time secure services that still preserves end-to-end trust?
- *Active Resilience and Evasive Safety.* How can PIE agents transition from passive "stop-on-fault" logic to active hazard avoidance? This requires research into detecting physical-world attacks, such as sensor spoofing or physical obstruction, and developing real-time planners that can identify "safe exit" trajectories to protect the agent and its cargo from ongoing threats.
- *Accountability and Forensic Logging.* Since PIE actions can be irreversible, auditability becomes essential. What is the right notion of a "black-box recorder" for embodied systems: what to log (e.g., inputs, model versions, safety checks), at what rate, and how to protect logs from tampering while respecting privacy?
- *Consent and Negotiation in Shared Spaces.* In the mall scenario, multiple stakeholders coexist: the user, bystanders, the facility operator, and vendors. How should consent be expressed and enforced when sensing inevitably captures bystanders? Can environments expose machine-readable privacy contracts that constrain what embodied agents may record and retain?
- *Trustworthy Human-PIE Symbiosis.* Trust also includes usability and interpretability. How should a PIE system communicate intent, uncertainty, and intervention rationale to users in a way that supports correct reliance (neither over-trust nor under-trust)? Can systems enforce calibrated autonomy via explicit "confidence-to-control" mappings?

6 Industry Inflection at CES 2026

As this visionary paper reaches completion in early 2026, the Consumer Electronics Show (CES) 2026 in Las Vegas, Nevada, provides a timely industrial inflection point. CES 2026 marks a decisive transition from generative AI toward PIE. The industry has moved beyond the "passive client" paradigm toward autonomous, embodied agents capable of reasoning about physical causality and acting in real time under stringent safety, energy, and timing constraints.

Across the exhibition floor, a consistent pattern emerges: leading companies are converging on solutions that explicitly bridge the long-standing systems gaps between cloud-scale reasoning and predictable, certifiable, and energy-efficient execution at the edge. Collectively, these efforts illustrate how PIE principles are beginning to materialize across hardware platforms, system software, and vertically integrated solutions.

NVIDIA. At CES 2026, NVIDIA positioned itself as the de facto operating system for Physical AI with the unveiling of Cosmos[②], a foundation model designed to reason over environments governed by real-world physics rather than purely statistical correlations. This direction closely aligns with PIE, where causal reasoning and local execution are essential. The Jetson T4000 module exemplifies this shift, delivering approximately four times the performance of previous generations within a 70-watt power envelope, enabling energy-constrained autonomy. When coupled with VLA models integrated into the Isaac robotics platform, NVIDIA enables robots to interpret human intent and execute physically grounded actions locally, substantially reducing cloud dependence and end-to-end latency.

AMD. AMD unveiled the Ryzen AI Embedded P100 and X100 Series[③], signaling a deliberate move beyond consumer AI acceleration toward systems designed for physical-world operation. Unlike general-purpose processors, these platforms are explicitly engineered for the demands articulated PIE research. From a PIE perspective, the P100 and X100 series directly address the timing predictability challenge by pairing Zen 5 CPU cores for deterministic control with an XDNA 2 neural processing unit capable of delivering up to 50 TOPS. Crucially, AMD introduced an ASIL-B-capable architecture, where ASIL (Automotive Safety Integrity Level) is an ISO 26262 safety classification, enabling these chips to manage safety-critical workloads in autonomous vehicles and industrial robots, where timing faults or execution jitter can lead to catastrophic physical outcomes.

Arm. Arm's launch of a dedicated Physical AI Business Unit[④] at CES 2026 signals a strategic consolidation of its automotive and robotics efforts around real-time, safety-critical intelligence. By standardizing compute architectures across servers, vehicles, and robots, Arm is constructing a seamless cloud-to-edge fabric that allows AI models to migrate without extensive software reengineering. From a PIE perspective, this approach directly addresses timing predictability and reflex safety, enabling real-time control loops to be embedded at the architectural level rather than imposed retroactively. The result is a hardware-software substrate explicitly designed for deterministic decision-making in physical systems.

Kodiak AI & Bosch. The partnership between Kodiak AI and Bosch[⑤] represents a pivotal step toward scalable, production-grade autonomy in long-haul trucking. By combining Kodiak's AI driving stack with Bosch's automotive-grade redundancy in steering and braking, the collaboration addresses one of PIE's central challenges: operating safely in an irreversible physical world. This joint platform ensures that even under sensor or subsystem failures, certified fallback behaviors remain available, preventing catastrophic outcomes. In doing so, the partnership closes the gap between intelligent decision-making and trustworthy physical actuation.

LG. LG's CLOiD robot[⑥] embodies the vision of a continuous, agentic computing system embedded within everyday environments. Featuring a torso with seven degrees of freedom, CLOiD can perform complex household tasks such as object manipulation and appliance interaction. From a PIE standpoint, CLOiD emphasizes affordance-based representations: rather than merely recognizing objects, the system understands how to grasp, balance, and interact with them under physical constraints. Integrated with LG's ThinQ ecosystem, CLOiD demonstrates how perception, reasoning, and actuation can be tightly coupled at the edge to enable practical "zero-labor" domestic automation.

Qualcomm. Qualcomm reaffirmed its leadership in software-defined vehicles at CES 2026 through the Snapdragon Cockpit Elite and Ride Elite platforms[⑦], now supporting agentic AI across multiple global automakers. These platforms exemplify PIE's need for cross-modal energy and resource management, as vehicles must balance power-intensive AI reasoning with mission-critical safety functions. Qualcomm's Digital Chassis acts as a centralized nervous system, dynamically allocating compute and thermal budgets to ensure that perception, planning, and control remain deterministic. This specialization avoids the nondeterministic throttling common in general-purpose hardware, making agentic mobility viable at scale.

[②] https://nvidianews.nvidia.com/news/nvidia-launches-cosmos-world-foundation-model-platform-to-accelerate-physical-ai-development, Jan. 2026.
[③] https://www.amd.com/en/newsroom/press-releases/2026-1-5-amd-introduces-ryzen-ai-embedded-processor-portfol.html, Jan. 2026.
[④] https://newsroom.arm.com/blog/the-next-platform-shift-physical-and-edge-ai-powered-by-arm, Jan. 2026.
[⑤] https://kodiak.ai/news/kodiak-bosch-scale-autonomous-trucking-hardware, Jan. 2026.
[⑥] https://www.lgcorp.com/media/release/29725/, Jan. 2026.
[⑦] https://www.qualcomm.com/news/releases/2026/01/leapmotor-and-qualcomm-debuts-world-s-first-automotive-central-c, Jan. 2026.

TIER IV (*Autoware*). A long-time leader in open-source autonomous driving, TIER IV showcased its end-to-end AI architecture targeting Level-4+ autonomy[⑧], signaling a deliberate move away from brittle, rule-based pipelines. By embracing monolithic end-to-end models, TIER IV directly addresses the semantic-to-physical translation problem that often limits real-world robustness. Through the Open AD Kit, the platform provides a tool-first execution environment in which open-source agents can be benchmarked, validated, and safely deployed across heterogeneous edge hardware, closely aligning with PIE's emphasis on reproducibility, validation, and predictable execution in safety-critical systems.

Samsung. While Samsung's Ballie robot has shifted toward internal research to address navigation complexity, the company's CES 2026 strategy emphasizes integrated home intelligence through its Bespoke AI[⑨] appliances and home hubs. These systems function as a passive yet anticipatory layer that mediates the home's physical infrastructure. In the context of PIE, Samsung focuses on intent hierarchies, where the system infers user needs, such as food management or hydration, and proactively adjusts the physical environment. This approach lays the groundwork for future embodied agents by first stabilizing and structuring the underlying physical context.

CES 2026 makes clear that Physical Intelligence on the Edge (PIE) has emerged as the unifying paradigm for deploying AI in the real world. Across sectors, leading systems now prioritize physically grounded reasoning, deterministic execution, and tight integration of perception, decision-making, and actuation under strict energy and safety constraints. This shift closes the long-standing gap between cloud-scale intelligence and edge deployment, underscoring that future autonomy will be defined less by model size and more by predictable, embodied, and trustworthy execution in the physical world.

7 Conclusions

Physical Intelligence on the Edge (PIE) represents a qualitative transition in computing: the edge is no longer merely a place to cache, filter, or accelerate inference, but the primary platform for closing the sense-think-act loop under the unforgiving constraints of the physical world. This transition shifts the systems objective from efficiently moving and processing "bits" to predictably and safely influencing "atoms". As illustrated through the hypothetical mall scenario, many enabling technologies already exist in isolation[51–53]. The central gap lies in integration: today's edge platforms can execute perception and inference, yet they do not reliably provide certifiable timing, safety envelopes, and robust autonomy when these components are composed into continuous closed-loop behavior.

With this background in place, we argued that PIE expands the edge agenda along five thrusts: embodied spatial reasoning, embodied temporal reasoning, edge-native customization, symbiosis, and sustainability. Across these thrusts, a recurring theme is that PIE demands cross-layer co-design. Perception, scheduling, verification, networking, and actuation must be engineered as a coherent system rather than as modular components optimized in isolation. In particular, temporal predictability and energy efficiency emerge as coupled constraints: unbounded latency, long-tail jitter, and thermal throttling are not performance nuisances but safety hazards. Likewise, privacy and trust become inseparable from correct operation: a system that cannot authenticate its inputs, enforce contextual consent, and audit its actions cannot be responsibly deployed at scale.

In terms of broader impacts, PIE provides a unifying systems research direction for next-generation autonomy in vehicles, robotics, smart infrastructure, and human-assistive technologies. If realized, PIE would enable embodied platforms that are not only capable, but also predictable, auditable, and sustainable, and systems that can earn trust in shared environments and deliver reliable assistance without constant cloud dependence. Achieving this vision will require new abstractions that make learning-based intelligence schedulable and monitorable, new verification interfaces that connect semantics to safe physical action, and new resource-management mechanisms that trade capability against energy and risk in principled ways. The payoff is substantial: an edge-native foundation for autonomous systems that can safely operate in the open world, at human timescales, and within real operational constraints.

Conflict of Interest Weisong Shi is an editorial board member for Journal of Computer Science

[⑧] https://www.prnewswire.com/news-releases/tier-iv-to-showcase-e2e-ai-for-level-4-autonomy-at-ces-2026-302652020.html, Jan. 2026.
[⑨] https://news.samsung.com/global/connected-comprehension-inside-samsungs-2026-ai-home, Jan. 2026.

and Technology and was not involved in the editorial review of this article. The authors declare that there are no other competing interests.

References

[1] Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 2016, 3(5): 637–646. DOI: 10.1109/JIOT.2016.2579198.

[2] Chiang M, Zhang T. Fog and IoT: An overview of research opportunities. *IEEE Internet of Things Journal*, 2016, 3(6): 854–864. DOI: 10.1109/JIOT.2016.2584538.

[3] Mao Y, You C, Zhang J, Huang K, Letaief K B. A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys & Tutorials*, 2017, 19(4): 2322–2358. DOI: 10.1109/COMST.2017.2745201.

[4] Satyanarayanan M. The emergence of edge computing. *Computer*, 2017, 50(1): 30–39. DOI: 10.1109/MC.2017.9.

[5] Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 2019, 107(8): 1738–1762. DOI: 10.1109/JPROC.2019.2918951.

[6] Cong J, Ghodrat M A, Gill M, Grigorian B, Reinman G. CHARM: A composable heterogeneous accelerator-rich microprocessor. In *Proc. the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design*, Jul. 2012, pp.379–384. DOI: 10.1145/2333660.2333747.

[7] Ji S, Chen X, Zhuang J, Zhang W, Yang Z, Schultz S, Song Y, Hu J, Jones A, Dong Z, Zhou P. ART: Customizing accelerators for DNN-enabled real-time safety-critical systems. In *Proc. the 2025 Great Lakes Symposium on VLSI*, Jul. 2025, pp.442–449. DOI: 10.1145/3716368.3735215.

[8] Ji S, Yang Z, Chen X, Zhang W, Zhuang J, Jones A, Dong Z, Zhou P. DERCA: DetERministic cycle-level accelerator on reconfigurable platforms in DNN-enabled real-time safety-critical systems. In *Proc. the 2025 IEEE Real-Time Systems Symposium*, Dec. 2025, pp.392–405. DOI: 10.1109/RTSS66672.2025.00039.

[9] Atat R, Liu L, Chen H, Wu J, Li H, Yi Y. Enabling cyber-physical communication in 5G cellular networks: Challenges, spatial spectrum sensing, and cyber-security. *IET Cyber-Physical Systems: Theory & Applications*, 2017, 2(1): 49–54. DOI: 10.1049/iet-cps.2017.0010.

[10] Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, Rafailov R, Foster E P, Sanketi P R, Vuong Q, Kollar T, Burchfiel B, Tedrake R, Sadigh D, Levine S, Liang P, Finn C. OpenVLA: An open-source vision-language-action model. In *Proc. the 8th Conference on Robot Learning*, Nov. 2024.

[11] Sumaiya, Jafarpourmarzouni R, Lu S, Dong Z. Enhancing real-time inference performance for time-critical software-defined vehicles. In *Proc. the 2024 IEEE International Conference on Mobility, Operations, Services and Technologies*, May 2024, pp.101–113. DOI: 10.1109/MOST60774.2024.00019.

[12] Sumaiya, Jafarpourmarzouni R, Luo Y, Lu S, Dong Z. Toward real-time and efficient perception workflows in software-defined vehicles. *IEEE Internet of Things Journal*, 2025, 12(6): 7240–7258. DOI: 10.1109/JIOT.2024.3492801.

[13] Roa M A, Berenson D, Huang W. Mobile manipulation: Toward smart manufacturing [TC spotlight]. *IEEE Robotics & Automation Magazine*, 2015, 22(4): 14–15. DOI: 10.1109/MRA.2015.2486583.

[14] Chevalier A, Copot C, De Keyser R, Hernandez A, Ionescu C. A multi agent system for precision agriculture. In *Handling Uncertainty and Networked Structure in Robot Control*, Busoniu L, Tamás L (eds.), Springer, 2015, pp.361–386. DOI: 10.1007/978-3-319-26327-4_15.

[15] Lee J, Kang S, Lee J, Shin D, Han D, Yoo H J. The hardware and algorithm co-design for energy-efficient DNN processor on edge/mobile devices. *IEEE Trans. Circuits and Systems I: Regular Papers*, 2020, 67(10): 3458–3470. DOI: 10.1109/TCSI.2020.3021397.

[16] Nowshin F, Dong Z, Yi Y. Memory-augmented autoencoder with reservoir computing for edge-based anomaly detection in autonomous systems. *IEEE Internet Computing*, 2025. DOI: 10.1109/MIC.2025.3594330.

[17] Jafarpourmarzouni R, Sumaiya F, Li R, Guan N, Wang G, Zhou P, Dong Z. Reaction latency analysis of message synchronization in edge-assisted autonomous driving. *ACM Trans. Embedded Computing Systems*, 2025. DOI: 10.1145/3736412.

[18] Liu T, Wang S, Li B, Dong Z, Wang G, Gong W, He T. Real-batch: Real-time adaptive batch processing for accurate object detection in autonomous driving. *IEEE Trans. Mobile Computing*, 2025. DOI: 10.1109/TMC.2025.3625072.

[19] Qian L, Luo Z, Du Y, Guo L. Cloud computing: An overview. In *Proc. the 1st International Conference on Cloud Computing*, Dec. 2009, pp.626–631. DOI: 10.1007/978-3-642-10665-1_63.

[20] Waldo J, Wyant G, Wollrath A, Kendall S. A note on distributed computing. In *Lecture Notes in Computer Science 1222*, Vitek J, Tschudin C (eds.), Springer, 1997, pp.49–64. DOI: 10.1007/3-540-62852-5_6.

[21] Birman K P. The process group approach to reliable distributed computing. *Communications of the ACM*, 1993, 36(12): 37–53. DOI: 10.1145/163298.163303.

[22] Wang Q, Yao Y, Shi W. Edge-assisted object perception for autonomous vehicles under challenging exposure and blur conditions. In *Proc. the 3rd IEEE International Conference on Mobility, Operations, Services and Technologies*, May 2025, pp.12–21. DOI: 10.1109/MOST65065.

2025.00011.

[23] Wu C, Gong Y, Liu L, Li M, Wu Y, Shen X, Li Z, Yuan G, Shi W, Wang Y. AyE-Edge: Automated deployment space search empowering accuracy yet efficient real-time object detection on the edge. In *Proc. the 43rd IEEE/ACM International Conference on Computer-Aided Design*, Oct. 2024, Article No. 178 DOI: 10.1145/3676536.3676655.

[24] Bhattacharjee A, Mahmood H, Lu S, Ammar N, Ganlath A, Shi W. Edge-assisted over-the-air software updates. In *Proc. the 9th IEEE International Conference on Collaboration and Internet Computing*, Nov. 2023, pp.18–27. DOI: 10.1109/CIC58953.2023.00013.

[25] Hao Y, Yang F, Fang N, Liu Y S. EMBOSR: Embodied spatial reasoning for enhanced situated question answering in 3D scenes. In *Proc. the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2024, pp.9811–9816. DOI: 10.1109/IROS58592.2024.10801720.

[26] Vila L. A survey on temporal reasoning in artificial intelligence. *AI Communications*, 1994, 7(1): 4–28. DOI: 10.3233/AIC-1994-7102.

[27] Zhang M, Cao J, Yang L, Zhang L, Sahni Y, Jiang S. ENTS: An edge-native task scheduling system for collaborative edge computing. In *Proc. the 7th IEEE/ACM Symposium on Edge Computing*, Dec. 2022, pp.149–161. DOI: 10.1109/SEC54971.2022.00019.

[28] Ajoudani A, Zanchettin A M, Ivaldi S, Albu-Schäffer A, Kosuge K, Khatib O. Progress and prospects of the human-robot collaboration. *Autonomous Robots*, 2018, 42(5): 957–975. DOI: 10.1007/s10514-017-9677-2.

[29] Li W, Yang T, Delicato F C, Pires P F, Tari Z, Khan S U, Zomaya A Y. On enabling sustainable edge computing with renewable energy resources. *IEEE Communications Magazine*, 2018, 56(5): 94–101. DOI: 10.1109/MCOM.2018.1700888.

[30] Duan J, Yu S, Tan HL, Zhu H, Tan C. A survey of embodied AI: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022, 6(2): 230–244. DOI: 10.1109/TETCI.2022.3141105.

[31] Liu T, Wang S, Dong Z, Li B, He T. From perception to computation: Revisiting delay optimization for connected autonomous vehicles. *ACM Computing Surveys*, 2025, 57(8): Article No. 200. DOI: 10.1145/3718361.

[32] Tian Z, Xia L, Shi W. EMATO: Energy-model-aware trajectory optimization for autonomous driving. In *Proc. the 2025 IEEE International Conference on Robotics and Automation*, May 2025, pp.9682–9688. DOI: 10.1109/ICRA55743.2025.11127833.

[33] Bouguettaya A, Kechida A, Taberkit A M. A survey on lightweight CNN-based object detection algorithms for platforms with limited computational resources. *International Journal of Informatics and Applied Mathematics*, 2019, 2(2): 28–44.

[34] Ouyang Y, Wu X, Yang M, Han R, Luo A, Liu C, Chen J, Guo Y. EdgeTail: Mitigating long-tail visual problems in continual learning at edge. *ACM Trans. Internet of Things*, 2025. DOI: 10.1145/3786774.

[35] Zhang Y, Kang B, Hooi B, Yan S, Feng J. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023, 45(9): 10795–10816. DOI: 10.1109/TPAMI.2023.3268118.

[36] Kaelbling L P, Littman M, Moore A W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996, 4: 237–285. DOI: 10.5555/1622737.1622748.

[37] Lakoff G. Cognitive semantics. In *Meaning and Mental Representations*, Eco U, Santambrogio M, Violi P (eds.), Indiana University Press, 1988, pp.119–154.

[38] Talmy L. Toward a cognitive semantics. *Concept Structuring Systems* , volume 1. MIT Press, 2000.

[39] Ball R, North C, Bowman D A. Move to improve: Promoting physical navigation to increase user performance with large displays. In *Proc. the 2007 SIGCHI Conference on Human Factors in Computing Systems*, Apr. 28-May 3, 2007, pp.191–200. DOI: 10.1145/1240624.1240656.

[40] Chen T, Yao Y, Hofstee H P, Shi W. Open-vocabulary object detection with driving-aware multi-scale feature fusion for autonomous driving. In *Proc. the 10th ACM/IEEE Symposium on Edge Computing*, Dec. 2025, Article No. 66. DOI: 10.1145/3769102.3774632.

[41] Liu L, Dong Z, Wang Y, Shi W. Prophet: Realizing a predictable real-time perception pipeline for autonomous vehicles. In *Proc. the 2022 IEEE Real-Time Systems Symposium*, Dec. 2022, pp.305–317. DOI: 10.1109/RTSS55097.2022.00034.

[42] Zhou Z, Li Y, Liu J, Li G. Equality constrained robust measurement fusion for adaptive Kalman-filter-based heterogeneous multi-sensor navigation. *IEEE Trans. Aerospace and Electronic Systems*, 2013, 49(4): 2146–2157. DOI: 10.1109/TAES.2013.6621807.

[43] Micucci D, Marchese F, Sorrenti D, Tisato F. A time-sensitive approach to mobile robot autonomous navigation. In *Proc. the 2004 International Conference on Computing, Communications and Control Technologies*, Aug. 2004, pp.377–382.

[44] Kaplan R, Friston K J. Planning and navigation as active inference. *Biological Cybernetics*, 2018, 112(4): 323–343. DOI: 10.1007/s00422-018-0753-2.

[45] Li R, Jiang X, Dong Z, Wu J M, Xue C J, Guan N. Worst-case latency analysis of message synchronization in ROS. In *Proc. the 2023 IEEE Real-Time Systems Symposium*, Dec. 2023, pp.185–197. DOI: 10.1109/RTSS59052.2023.00025.

[46] Xu Y, Musgrave Z, Noble B, Bailey M. Bobtail: Avoiding long tails in the cloud. In *Proc. the 10th USENIX Symposium on Networked Systems Design and Implementation*, Apr. 2013, pp.329–341. DOI: 10.5555/2482626.2482658.

[47] Marti P, Fuertes J M, Fohler G, Ramamritham K. Jitter compensation for real-time control systems. In *Proc. the*

22nd IEEE Real-Time Systems Symposium, Dec. 2001, pp.39–48. DOI: 10.1109/REAL.2001.990594.

[48] Tong L, Li Y, Gao W. A hierarchical edge cloud architecture for mobile computing. In *Proc. the 35th Annual IEEE International Conference on Computer Communications*, Apr. 2016. DOI: 10.1109/INFOCOM.2016.7524340.

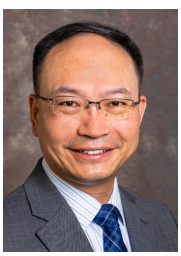
[49] Herbert S, Marculescu D. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In *Proc. the 2007 International Symposium on Low Power Electronics and Design*, Aug. 2007, pp.38–43. DOI: 10.1145/1283780.1283790.

[50] Choudhary T, Mishra V, Goswami A, Sarangapani J. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 2020, 53(7): 5113–5155. DOI: 10.1007/s10462-020-09816-7.

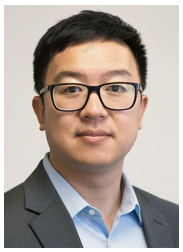
[51] Luo Q, Luan T, Shi W, Fan P. Deep reinforcement learning based computation offloading and trajectory planning for multi-UAV cooperative target search. *IEEE Journal on Selected Areas in Communications*, 2023, 41(2): 504–520. DOI: 10.1109/JSAC.2022.3228558.

[52] Wu Z, Wang S, Bao Y, Shi W. Tentacles: A middleware with multi-network communication reliability for vehicle-infrastructure cooperative autonomous driving. In *Proc. the 100th IEEE Vehicular Technology Conference*, Oct. 2024. DOI: 10.1109/VTC2024-Fall63153.2024.10758047.

[53] Luo Q, Luan T, Shi W, Fan P. Edge computing enabled energy-efficient multi-UAV cooperative target search. *IEEE Trans. Vehicular Technology*, 2023, 72(6): 7757–7771. DOI: 10.1109/TVT.2023.3238040.



Weisong Shi is an Alumni Distinguished Professor and Chair of the Department of Computer and Information Sciences at the University of Delaware (UD), Newark, where he leads the Connected and Autonomous Research (CAR) Laboratory. He is an internationally renowned expert in edge computing, autonomous driving, and connected health. He is the Editor-in-Chief of IEEE Internet Computing Magazine and Elsevier Smart Health. He is the founding steering committee chair of three conferences, including the ACM/IEEE Symposium on Edge Computing (SEC), the IEEE/ACM International Conference on Connected Health (CHASE), and the IEEE International Conference on Mobility (MOST).



Zheng Dong is an associate professor in the Department of Computer Science at Wayne State University, Detroit. He received his B.S. degree from Wuhan University, Wuhan, in 2007, his M.S. degree from the University of Science and Technology of China, Hefei, in 2011, and his Ph.D. degree from The University of Texas at Dallas, Richardson, in 2019. His research interests include real-time embedded AI systems, edge computing, and connected and autonomous driving systems. He received the Outstanding Paper Award at the 38th IEEE Real-Time Systems Symposium (RTSS) and a Best Paper nomination at the 23rd IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA). He serves as a Steering Committee Co-Chair of the IEEE Workshop on Physical Intelligence: Systems and Applications (PISA) and is a recipient of the NSF CAREER Award and the NSF CRII Award.

Peipei Zhou is currently an assistant professor at the School of Engineering, Brown University, Providence. She received her Ph.D. degree in computer science and her M.S. degree in electrical and computer engineering from University of California, Los Angeles, in 2019 and 2014, respectively, and her B.S. degree in electrical and computer engineering from Southeast University, Nanjing, in 2012. Her research investigates architecture, programming abstraction, and design automation tools for reconfigurable computing and heterogeneous computing. She has published 40 papers in IEEE/ACM computer system and design automation conferences and journals. Her work has won the 2025 IEEE/ACM ICCAD 10-Year Retrospective Most Influential Paper Award and the 2019 IEEE TCAD Donald O. Pederson Best Paper Award. Other awards include the 2024 ACM/IEEE IGSC Best Viewpoint Paper, the 2025 ACM/SIGDA FPGA Best Paper Nominee, the 2018 IEEE ISPASS Best Paper Nominee, and the 2018 IEEE/ACM ICCAD Best Paper Nominee.