[Emerging Ideas] Artificial Tripartite Intelligence: A Bio-Inspired, Sensor-First Architecture for Physical AI

You Rim Choi*
Seoul National University
yrchoi@snu.ac.kr

Subeom Park*
Seoul National University
sbpark7@snu.ac.kr

Hyung-Sin Kim
Seoul National University
hyungkim@snu.ac.kr

Abstract

As AI shifts from data centers to robots and wearables, the scaling recipe of ever-larger models becomes insufficient. Embodied systems operate under tight latency and power constraints, relying on controllable sensors in dynamic environments. In these settings, robustness requires not just enlarging a remote "cerebrum" but improving the input itself. We introduce Artificial Tripartite Intelligence (ATI), a bio-inspired architecture for physical-world AI. ATI distributes intelligence across four roles: a millisecond-scale Brainstem (device) for safety, a sensor-rate Cerebellum (device) for sensor calibration, a fast Cerebrum-F (device) for initial perception, and a deep Cerebrum-D (edge/cloud) for high-cost reasoning. By elevating the Brainstem and Cerebellum to proactively shape input, ATI handles most cases locally, reducing latency, offload fraction, and reliance on compressed giant models. We evaluate ATI on a mobile camera prototype in dynamic lighting conditions. Results show that active stabilization doubles downstream accuracy (52% vs. 26%) and reduces cloud offloading by ~37%. This demonstrates that for resilient embodied intelligence, sensor-first architecture is as critical as model scale.

ACM Reference Format:
You Rim Choi*, Subeom Park*, and Hyung-Sin Kim. 2025. [Emerging Ideas] Artificial Tripartite Intelligence: A Bio-Inspired, Sensor-First Architecture for Physical AI. In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Over the past decade, AI has advanced chiefly by scaling datasets and models in the cloud [3, 43, 50]. However, as AI enters the era of *physical, embodied AI* [19, 33, 35], this recipe breaks down. Physical inputs arrive through *device-anchored, controllable sensors* (e.g., cameras, microphones, IMUs) that operate in dynamic environments and share tight power, latency, and privacy budgets with on-device and edge computation. Yet most responses to this shift remain *computation-centric*: from domain adaptation to model compression and offloading, the prevailing assumption still treats sensors as static front-ends and concentrates almost all adaptivity inside a single model stack (Figure 1(a)).

Biological perception offers a sharp contrast. Human sensing is inherently *adaptive*: mechanisms such as pupil reflexes and retinal gain control continuously reshape raw signals *before* they reach the cortex (Table 1). Conversely, artificial sensors typically operate in a static regime, pushing the burden of handling noise, glare, and motion almost entirely onto the downstream model. A growing body of work on *adaptive sensing* begins to treat sensing as control. For example, recent systems [2, 31] tune camera ISO or shutter speed to adjust brightness before feeding images to a vision model. However, current methods either rely on simple heuristics isolated from inference, or tightly couple a single sensor to a heavy model that is repeatedly queried to maximize its own confidence. Consequently, they remain *cortex-centric*, stopping short of architecting a unified intelligence stack in which sensors and "brains" are co-designed partners (Figure 1(b)).

In this paper, we argue that sensor-brain co-design for physical AI is fundamentally a *systems architecture* problem. Just as the von Neumann architecture [51] brought an inflection point via a *systems contract* for programmable computation, instead of adding more vacuum tubes and faster arithmetic units, physical AI needs a contract that recognizes two inherent time scales: millisecond-level safety-critical reflexes and slower, more complex reasoning over long-tail and multimodal situations. Compressing a gigantic cortex onto a device often leaves it too weak for difficult cases, while relying solely on the cloud makes fast reflexes brittle. To bridge this gap, we propose **Artificial Tripartite Intelligence (ATI)**, an architectural contract that starts intelligence at

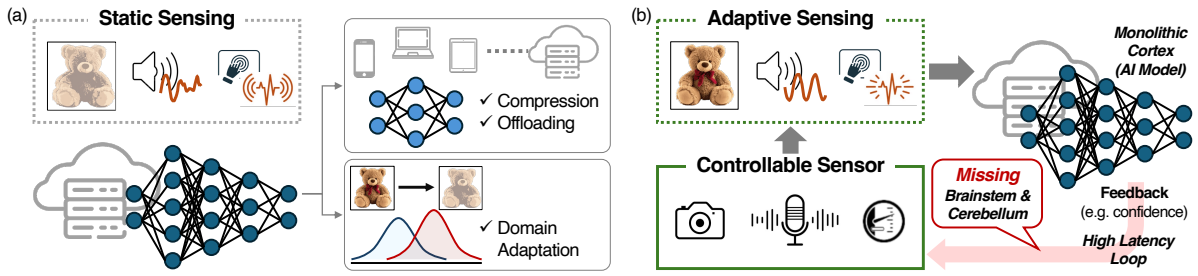

Figure 1: Limitations of prevailing paradigms for physical AI. (a) Computation-centric approach: The focus is solely on optimizing the model stack via compression, offloading, or domain adaptation to handle resource constraints and distribution shifts, while sensors are treated as static, passive data pipes. (b) Cortex-centric adaptive sensing: A monolithic cortex (representing a large perception or reasoning model) directly attempts to control the sensor based on high-level inference confidence. This lacks the crucial fast reflex loops (brainstem-like) and predictive calibration planes (cerebellum-like) found in biological systems, resulting in brittle or inefficient operation.

Table 1: Adaptive human sensing versus conventional artificial sensors across four modalities.

Modality	Human sensor (adaptive)	Artificial sensor (static)
Vision	Pupil dynamically controls light intake. Photoreceptors adapt to luminance over seconds. Fast saccades redirect the fovea before cortical processing.	*Camera* (controls: exposure, ISO, frame rate, focus). Frame rate and exposure/ISO set in coarse steps, rarely changed. Saturation and motion blur corrected mainly after capture.
Hearing	Active gain control via outer hair cells over a wide dynamic range. Efferent feedback protects against loud sounds. Sub-millisecond binaural timing enables precise localization.	*MEMS microphone* (controls: gain/AGC, sampling rate, beam pattern). AGC, sampling rate, and bit depth fixed at design time. 3D localization relies on multi-mic arrays and heavy DSP.
Touch	Dense, compliant skin with high spatial and temporal acuity. Reflex arcs withdraw from painful stimuli before awareness. Active exploration (rubbing) increases signal-to-noise ratio.	*Tactile array / pressure sensor* (controls: readout freq, active ROI). Sparse capacitive or piezoresistive taxel arrays on rigid surfaces. Thresholds and sampling rates fixed; protection left to high-level control.
Proprioception	Vestibular and proprioceptive organs track body pose and acceleration. Rapid reflexes stabilize posture and gait without cortical involvement.	*IMU* (controls: measurement range, sampling rate, filter parameters). Accel/gyro data sampled at fixed rates with offline bias calibration. Ranges and filters rarely change after deployment.

the sensor, budgets latency through well-defined lanes, and places computation across device, edge, and cloud according to explicit policy.

What is ATI? ATI is a **sensor-oriented intelligence layer** running on top of existing operating systems (OSes) and middleware [58]. Inspired by the biological brain (brainstem, cerebellum, and cerebrum), it functionally decomposes the intelligence stack into four cooperating roles:

- **Brainstem (reflex).** An ultra-fast plane, tightly coupled to sensors and actuators, that enforces non-bypassable safety and signal-integrity invariants at the millisecond scale.
- **Cerebellum (calibration).** An adaptive plane running at sensor rate that uses high-level model feedback (e.g., confidence signals) to learn a lightweight internal model, and then autonomously tunes sensor parameters (e.g., exposure, gain, ROI) so that the input stream remains "model-friendly" without frequent cortical queries.
- **Cerebrum-F (fast perception).** A lightweight on-device cortical plane that provides the first useful interpretation and routing for common cases, so that routine decisions remain responsive even without reliable connectivity.
- **Cerebrum-D (deep reasoning).** A deep cortical plane residing on the edge or cloud for long-tail and multimodal reasoning, invoked only when local confidence is low or the task demands richer semantic insight.

ATI elevates Brainstem and Cerebellum to *first-class citizens*. Device-side intelligence owns what must be fast and always available (Brainstem, Cerebellum, and Cerebrum-F) and escalates to Cerebrum-D only when the expected gain justifies the added latency, energy, and bandwidth cost. This cooperation allows a lightweight Cerebrum-F to handle most dynamic real-world scenarios effectively, reducing reliance on a remote "gigantic cerebrum."

Why *MobiSys*, and Why Now? The *MobiSys* community has long mastered efficient sensing under tight constraints [10, 39, 55]. Meanwhile, web-scale AI has built powerful "cerebrums" [12, 43] yet left them largely decoupled from real-time sensor control. As these two lineages converge in the era of embodied AI, ATI offers a principled architectural blueprint for this union. By transforming isolated algorithms

into a unified system, ATI takes the critical next step toward robust, sensor-integrated intelligence in the physical world.

Contributions. In this Emerging Ideas paper, we do not present a finished system but instead articulate a bio-inspired, systems-grounded architecture for physical-world AI, arguing that architecture, not just scale, is the critical next lever.

- **Architecture over algorithms.** We reframe adaptive sensing as a systems problem, moving from cortex-centric model scaling to a sensor-oriented architecture with explicit roles and latency lanes.
- **The ATI contract.** We define a portable intelligence layer with four roles (Brainstem, Cerebellum, Cerebrum-F, and Cerebrum-D) and minimal interfaces for sensor–model cooperation across modalities.
- **Prototype and outlook.** We demonstrate a vision-centric ATI prototype on a mobile system, showing how active sensor tuning in a modular architecture unlocks robust perception even with smaller local models, and we outline extensions to other sensing modalities.

2 Related Work

ATI intersects several threads in embodied AI and mobile systems: foundation models, resource-aware mobile/edge execution, adaptive sensing, and robotics middleware.

2.1 Foundation Models for Embodied AI

At the top of the physical-AI stack, embodied foundation models act as **a "Gigantic Cerebrum"** for high-level perception, reasoning, and action. CLIP [43] demonstrated large-scale vision–language alignment, and PaLM-E [12] extended this capability to embodied settings by jointly encoding images, robot state, and language. Vision–language–action (VLA) models further integrate perception and control: RT-2 [59] maps visual observations and language instructions directly to robot actions, and recent Gemini-based VLA systems [48] show that such agents can operate across diverse robot platforms. These systems deliver strong deep perception and semantic reasoning, which ATI positions as Cerebrum-D, but they generally assume stable, preconditioned inputs and do not specify how sensing, reflexes, or predictive calibration interact with them or how responsibilities should be allocated across device, edge, and cloud.

2.2 Computation-Centric Edge AI

Mobile and edge sensing systems focus primarily on *where and how* inference should run under tight resource and latency constraints. Early work established the foundations of hierarchical sensing [34, 39] and computation offloading [10, 45] to manage energy and latency. With the rise of deep neural networks (DNNs), the focus shifted to optimizing the *model stack* itself. Approaches range from model compression and dynamic scaling [14, 17] to scheduling [23, 54] and concurrent execution [16, 55]. Recent web-based approaches further democratize access via just-in-time kernel generation [24]. Collectively, these systems treat computation placement as a first-class concern but assume a fixed, passive sensing pipeline. ATI differs by assigning explicit roles for *sensor regulation* through Brainstem and Cerebellum, integrating sensing into the intelligence loop rather than optimizing only the cortical side.

2.3 Adaptive Sensing and Sensor Control

A complementary line of work treats parts of the sensing stack itself as a control problem. Classical systems adjust sensor parameters using hand-crafted heuristics, e.g., keeping image brightness near a target by tuning the exposure or gain of a camera. More recent systems adopt learning-based *adaptive sensing*. Several methods use deep reinforcement learning to control camera exposure and gain as part of a closed loop, either to improve general detection quality [31] or to stabilize visual odometry under challenging lighting [57]. Lens [2] introduces a lightweight "vision test" module that evaluates candidate camera settings (ISO, shutter speed, and aperture) using the target model's confidence as a proxy for image quality, then selects the best setting to improve accuracy without modifying the model itself. Other frameworks control pan-tilt-zoom (PTZ) motion to keep low-confidence targets within a high-quality field of view [52].

These approaches show that *sensing for the model* can matter as much as the model itself. However, they operate at the level of a single camera coupled to a single model, and they do not specify where sensor controllers should live in a device-edge-cloud stack, how they relate to safety-critical reflexes versus deeper reasoning, or how they should coordinate with routing between lightweight and heavy models. ATI lifts these ideas into an *architectural* contract by assigning sensor control to Brainstem and Cerebellum and by defining how these planes interact with fast and deep cortical planes (Cerebrum-F/D) within a unified runtime.

2.4 Robotics Middleware and Cloud Robotics

Robotics middleware such as Robot Operating System (ROS) and ROS 2 provides the standard abstraction for modular robot software through nodes, topics, services, and executors [36, 42]. Orocos [4] extends this model with hard real-time control primitives. More recently, FogROS2 [20] integrates ROS 2 with cloud and fog computing, enabling developers to launch ROS nodes in the cloud with minimal

code changes. Fault-tolerant extensions further address the unreliability of cloud resources [5, 6], illustrating that even well-engineered clouds fail and that robotics stacks require multi-path resilience. ATI complements rather than replaces these platforms. While ROS and FogROS2 provide the *mechanism* for scheduling and connectivity, ATI defines the *architectural policy*. It specifies *what* the components should be: which roles must enforce local safety invariants (Brainstem, Cerebellum, Cerebrum-F), which can be elastically offloaded (Cerebrum-D), and how responsibilities are partitioned across the device–edge–cloud hierarchy.

3 Artificial Tripartite Intelligence (ATI)

This section instantiates Artificial Tripartite Intelligence (ATI) as a concrete intelligence layer: how its four roles are distributed across device–edge–cloud resources, which "knobs" each role controls, which APIs connect them, and why this structure enables compressing the main on-device model while improving latency, robustness, and energy efficiency. We also describe how each role is trained. We present ATI primarily through visual perception since both camera control and visual neuroscience have been extensively studied, which lets us give concrete, grounded examples. The same principles, however, generalize to other modalities beyond vision.

3.1 Biological Insights

We first outline the design philosophy that motivates this bio-inspired, sensor-first decomposition. In vision, speech, and NLP, progress has followed a familiar trajectory: start with modular pipelines, then collapse them into end-to-end DNNs once data and compute scale up [29]. This recipe works well when the world presents pre-captured pixels or tokens and the primary trade-off is compute versus accuracy. In embodied settings, however, *this recipe breaks*. Some failures are created at capture, such as saturation, banding, and motion smear, and *cannot be learned away later*. A single ever-larger model cannot recover information that never reached the sensor in usable form, nor can it guarantee the first safe action within millisecond deadlines.

Biological vision addresses this physical reality with a layered control structure rather than a single processor. The **Brainstem** hosts fast visual reflexes that protect signal integrity and safety. The **Cerebellum** performs predictive calibration, stabilizing sensing and motion at roughly video rate. **Cortical pathways** in the **Cerebrum** then provide a rapid feed-forward gist and, when needed, slower recurrent refinement. We translate these biological motifs into a concrete systems contract: four roles—Brainstem, Cerebellum, Cerebrum-F, and Cerebrum-D—with explicit latency lanes, control knobs, and training objectives.

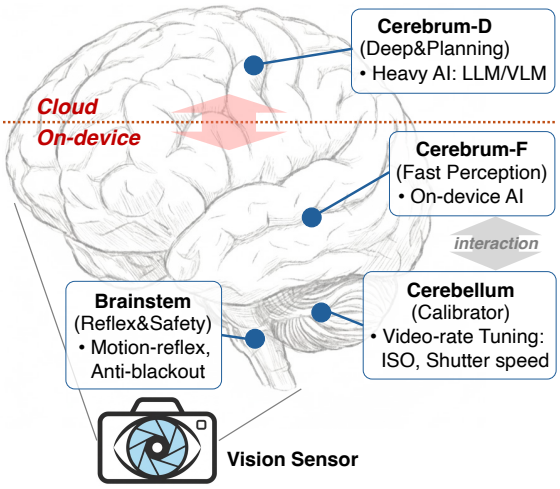

Figure 2: Artificial Tripartite Intelligence (ATI) for a camera-based system. The Brainstem and Cerebellum provide low-latency sensor control on the device, while Cerebrum-F and Cerebrum-D handle perception and reasoning at different time scales.

3.2 Sensor-First Roles and Placement

Each role in ATI (Figure 2) is defined by its biological inspiration, its system responsibility, and its hardware placement.

3.2.1 *Brainstem (L1: Reflex & Safety, device-only).*

Bio-Inspiration. Analogous to the vestibulo-ocular reflex (VOR) [8, 21], which stabilizes retinal images within milliseconds, and the pupillary light reflex [11, 56]. These circuits operate *before* cortical processing to protect the quality of the input: they prevent irreversible failures such as motion smear or saturation that no downstream model can recover.

ATI Implementation. A small, non-bypassable reflex plane tightly coupled to the camera driver and local actuators. It resides strictly on the device to satisfy a sub-frame budget (<10 ms). It enforces hard invariants, such as anti-flicker guards (rejecting shutter speeds that beat with lighting), anti-saturation clamps (preventing clipping), and thermal/laser safety. Brainstem acts as a hard gatekeeper for actuation: downstream models cannot force the sensor into states that violate these invariants.

3.2.2 *Cerebellum (L2: Calibration, device-only).*

Bio-Inspiration. Unlike simple reflexes, the biological cerebellum coordinates complex loops such as smooth pursuit

(stabilizing tracking) [32] and saccade adaptation (calibrating gaze shifts) [18]. It achieves this using learned internal models [22]—predictive simulators that estimate how sensor actions (e.g., exposure change) impact the quality and efficacy of downstream perception. This allows the system to reshape sensing proactively, stabilizing input quality against motion and lighting changes so that downstream cortical inference is simpler and more robust.

ATI Implementation. An adaptive plane operating on a sensor-synchronous timescale on the device (e.g., per-frame callbacks). It utilizes lightweight self-signals such as blur score, saturation ratio, banding energy, and SNR to continuously tune capture parameters: exposure, analog gain, HDR bracketing, tone mapping, ROI, and micro-gimbal actuation. Critically, it internalizes high-level feedback: while the Cerebrum defines the information goal ("task-directed sensing" [28, 53]), the Cerebellum learns the sensor policy that best supports it. For instance, it can learn to hold exposure on a specific face ROI identified by the cortex, maintaining a "model-friendly" signal autonomously as lighting shifts, without stalling for heavy cortical updates.

3.2.3 *Cerebrum-F (L3: Fast Perception, device-resident).*

Bio-Inspiration. Corresponds to the rapid feed-forward sweep ("vision without awareness") [49] and the dorsal stream ("how/where" pathway) [38, 46]. Biology separates this fast pathway (<150 ms) to support visually guided action and coarse categorization even before conscious recognition is complete. It provides the *gist* of the scene to ensure responsiveness.

ATI Implementation. A lightweight perception head running on the device accelerator NPU/GPU. Within a strict latency budget (tens of milliseconds), it produces presence decisions, ROIs, short-term tracks, and coarse class labels for common cases. Crucially, it hosts the Router: it computes a calibrated uncertainty score to decide when the current input is too ambiguous or "long-tail" to handle locally, triggering an escalation to L4. This layer ensures that basic interactivity (e.g., tracking a person) continues uninterrupted even if the network fails.

3.2.4 *Cerebrum-D (L4: Deep Reasoning, edge/cloud).*

Bio-Inspiration. Analogous to the slow recurrent loops of "conscious vision" [27] and the ventral stream ("what" pathway) [37]. Just as the brain invests time in recurrent processing to resolve ambiguity, integrate context, and support semantic awareness, L4 provides the depth required for complex understanding, trading latency for precision.

ATI Implementation. A deep foundation model (e.g., VLM, VLA) hosted on edge or cloud nodes. It handles rare, open-set, or logic-heavy queries: fine-grained identification, small-font OCR, 6-DoF pose estimation, and complex reasoning. Because of network variability, L4 is treated as advisory and deadline-bounded. Its output is an asynchronous "hint"; if a result arrives after the task's deadline, Cerebrum-F treats it as expired. Remote intelligence refines the local model's decisions but never blocks the local safety or control loops.

3.3 Minimal Interfaces: The Contract

To keep ATI portable and debuggable, each role interacts only through small, auditable interfaces. Conceptually, there are three primary APIs.

3.3.1 *Sensor Control Interface (L1/L2 → Sensor).*

Brainstem and Cerebellum share a low-level interface to the sensor and local actuators, with layered authority: Brainstem defines hard safety envelopes, and Cerebellum operates only within those bounds. Table 2 lists representative primitives exposed by this interface.

Table 2: Example sensor control APIs exposed by ATI for camera and gimbal control.

Function	Description
`set_shutter(step)`	Exposure duration (discrete shutter step)
`set_gain(dB)`	Analog signal amplification (in dB)
`set_hdr(mode)`	HDR / WDR bracketing strategy
`set_roi(box)`	Active pixel region read-out (ROI)
`slew_gimbal(Δp, Δt)`	Pan/tilt actuation velocity

The **Brainstem** uses these primitives to enforce caps and safe sets (e.g., rejecting shutter speeds that cause flicker; clamping exposure to prevent thermal damage). Within those safety envelopes, the **Cerebellum** schedules fine-grained per-frame adjustments at video rate to keep inputs within "model-friendly" regimes.

3.3.2 *Sensor-Cortex Loop (L2 ↔ L3).*

This bidirectional interface couples active sensing with fast perception.

- **Downstream (L2 → L3):** For each frame, L2 attaches a Quality Vector (QV)—metrics like blur score, saturation ratio, and estimated lux. This allows L3 to distinguish between "empty scene" and "blinded sensor", conditioning its uncertainty.
- **Upstream (L3 → L2):** L3 provides high-level task feedback, such as detection confidence, target ROI, or a "re-sample" request.

Internalization: Initially, L2 relies on this feedback to tune parameters (e.g., increase gain when confidence drops). Over time, L2 uses these signals to train its internal model, eventually learning to maintain optimal capture settings autonomously (like a learned reflex) without waiting for L3's feedback loop.

3.3.3 *Escalation Interface (L3 → L4).*

When L3 is uncertain even with optimized sensing, it issues targeted requests to the cloud via `escalate(payload, budget)`. The `payload` minimizes bandwidth (e.g., a cropped ROI), and the response is advisory. To prevent stale updates, the response carries the originating frame's timestamp, allowing L3 to discard results that arrive after the scene context has shifted. Crucially, L4 interacts only with L3; it does not control the sensor directly. L3 integrates L4's semantic insight (e.g., "this is a small text") and may translate it into new feedback for L2 (e.g., "focus on this region"), preserving the latency hierarchy.

3.4 Quality-Aware Routing Under Latency Constraints

The Router in Cerebrum-F answers a critical question per frame: *Is it worth asking Cerebrum-D for help, given my uncertainty, the input quality, and the current resource budget?* The router operates on four inputs: (1) **calibrated uncertainty** $u(x)$ from the fast head (e.g., entropy or conformal score), (2) **quality vector (QV)** from the Cerebellum (blur, saturation, low-lux flags), (3) **resource state** (current network round-trip time T_{RTT} and device energy headroom), and (4) **task metadata** (strict deadline $T_{deadline}$, e.g., the inter-frame interval or application-defined latency limit, and cost of error). Escalation to L4 is triggered only if all the following hold:

- **High Uncertainty:** The fast head is genuinely unsure ($u(x) > \tau_{task}$).
- **Sufficient Quality:** The QV confirms the sensor is not blinded. If the image is pitch black or severely smeared despite L2's best efforts, L4 will likely fail too. In this case, the router aborts escalation and instead requests a *resample* from L2.
- **Feasible Deadline:** The remaining time budget covers the full offloading cycle. Specifically, the expected arrival time—calculated as current time (T_{now}) plus estimated network round-trip (T_{RTT}) and remote inference latency (T_{inf_L4})—must precede the task's strict deadline ($T_{deadline}$):

$$T_{now} + T_{RTT} + T_{inf\_L4} \leq T_{deadline}$$

- **Net Benefit:** The trade-off is positive. We model this as a simple utility check: the expected reduction in uncertainty (proportional to $u(x)$) must exceed the estimated transmission energy cost weighted by current battery status.

This logic transforms the classic offloading problem into a quality-gated function placement problem. Unlike traditional systems that blindly offload hard frames, ATI filters out "hopeless" inputs locally. Crucially, the temporal logic ensures non-blocking safety: local control loops never stall waiting for L4. L4's output is treated as a future refinement; if it arrives late, it is discarded, and the system proceeds with the L3 decision to maintain the control frequency.

3.5 Role-Aligned Learning Strategy

ATI rejects the monolithic end-to-end training paradigm. Instead, its training strategy mirrors the runtime decomposition: each module is trained via objectives strictly aligned with its architectural role.

- **Brainstem (L1): Configured, Not Learned.** Safety cannot be guessed. Brainstem rules are engineered constraints (e.g., max thermal load, anti-flicker lookup tables) validated via regression tests. They are treated as firmware: deterministic and immutable during runtime to guarantee the "hard gatekeeper" function.
- **Cerebellum (L2): Task-Driven Policy.** Calibration is formulated as a control problem driven by downstream performance. Offline, sweep traces are collected to fit a policy that maximizes image quality metrics. Online, L2 performs *bounded* exploration inside the Brainstem's safety envelope. Crucially, *feedback* from L3 (e.g., detection confidence) is utilized as a supervision signal to refine the internal model, allowing the policy to map scene conditions to optimal sensor states that maximize task accuracy.
- **Cerebrum-F (L3): Distillation on Stabilized Inputs.** Because L2 guarantees a "model-friendly" input distribution (e.g., well-exposed, minimal blur), L3 does not need to learn robustness to every possible physical artifact. Training relies on *knowledge distillation* from L4. Hard examples identified by the Router are escalated to L4, labeled, and subsequently used to fine-tune L3. This enables a compact model to achieve high accuracy on the specific domain.
- **Cerebrum-D (L4): Generalist Teacher.** This role adopts a pre-trained foundation model, frozen or lightly adapted via prompting, rather than requiring training from scratch. This generalist serves a dual purpose: resolving long-tail queries at runtime and providing pseudo-labels to supervise L3 and L2 offline.
- **Router: Data-Driven Calibration.** The router relies on offline execution-based *calibration* (tuples of uncertainty, QV, and task outcome). Rather than enforcing a specific learning architecture, the decision logic comes from mapping the "uncertainty-quality" space to empirical failure rates. This process yields data-driven decision boundaries—ranging from fitted risk estimators to simple calibrated thresholds—that satisfy the benefit-cost logic (§3.4).

Why this matters. This decomposition yields two system-level benefits. (1) Data Efficiency: L3 is trained on a cleaner, lower-variance distribution thanks to L2's active stabilization, requiring significantly less data than end-to-end models that must learn to "see through" bad sensors. (2) Compute Efficiency: L1/L2 act as *pre-semantic filters*, dropping uninformative or irrecoverable frames before they waste NPU or network cycles.

4 ATI Prototype: Classifying on the Rail

To demonstrate the feasibility of ATI, we instantiated all four roles on a commodity mobile platform. We detail the hardware setup and the algorithms used for each role.

4.1 Task Scenario

We mount a Galaxy S25 smartphone on a toy car driving along a closed track. The phone runs a custom Android application that continuously captures RGB frames and performs object classification on the live stream.

Task Definition. The task is a *lap-based* classification challenge. At any given time, a single physical object is placed along the track. As the car completes a lap, it encounters the object intermittently. The system's goal is to correctly identify the class of the object within the duration of the lap. This requires the vision stack to be robust not just to static noise, but to the transient motion blur and varying viewing angles induced by the vehicle's movement.

Environment Control. We control the illumination at two levels, "bright" and "dark", and alternate these conditions during evaluation. This setup creates a challenging dynamic range where the system must trade off motion blur (caused by long exposure) against noise (caused by high ISO). On the Galaxy S25, we expose these two primary control knobs—ISO gain and shutter speed—to the ATI intelligence layer.

Inference Stack. On the inference side, the app runs an on-device classifier (Cerebrum-F, L3) using MobileNetV2 [44] pre-trained on ImageNet [26], or alternatively offloads frames to a large remote model (Cerebrum-D, L4) via the Gemini API [9]. This yields a concrete instance of ATI's split cortex operating on a highly dynamic camera stream.

4.2 L1: Rule-Based Sensor Control

The Brainstem (L1) acts as a hard gatekeeper at the sensor layer. Its primary objective is to prevent catastrophic image degradation—specifically severe motion blur, heavy under-exposure, or signal saturation—before any frame is captured. Consistent with the ATI design, L1 enforces these invariants by mapping continuous sensor readings into semantic states and applying a strict Safety Envelope.

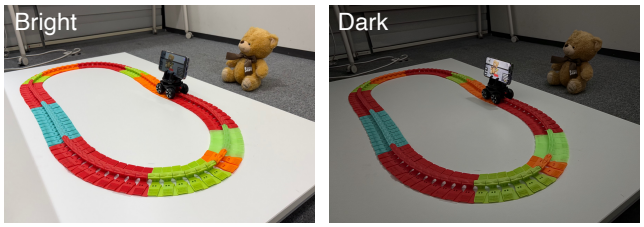

Figure 3: Experimental setup for the ATI prototype. A Galaxy S25 is mounted on a small toy vehicle that drives along a closed plastic track. We control ambient illumination at two levels: Bright (left) and Dark (right). The car completes one lap in approximately 3 s, creating motion and lighting changes while the camera captures frames for downstream classification.

- **Motion-based Shutter Limits.** L1 aggregates accelerometer and gyroscope magnitudes to classify motion into five levels: {STATIC, SLOW, NORMAL, FAST, SHAKE}. In high-motion states (e.g., FAST), L1 enforces a strict *maximum exposure time* to prevent irreversible motion blur. Crucially, this constraint forces the system to *compensate with higher ISO* to maintain image brightness. Conversely, in STATIC states, L1 relaxes the shutter limit, allowing longer exposures to *reduce ISO noise*.
- **Illuminance-based ISO Floors.** Ambient light is discretized into five context levels ranging from DARK (< 25 lux) to OUTDOOR (> 300 lux). In DARK or DIM environments, L1 enforces a *minimum ISO floor*. This prevents the exposure algorithm from dropping sensitivity too low, which would cause heavy under-exposure when the shutter speed is already constrained by motion.

Discretized Action Space and Safety Envelope. To minimize the computational overhead for the downstream L2 learner, we discretize the camera control knobs into a finite set of ISO and shutter speed steps. Given the current state tuple (s_{motion}, s_{light}), L1 calculates a valid action set $\mathcal{S}_{safe}$:

$$\mathcal{S}_{safe} = \{(t, g) \mid t \leq T_{cap}(s_{motion}),\ g \geq G_{floor}(s_{light})\}$$

where T_{cap} defines the motion-dependent shutter limit and G_{floor} defines the lux-dependent ISO floor. Any control command falling outside $\mathcal{S}_{safe}$ is clamped to the nearest valid boundary. This ensures the camera remains in a safe operating region—free from extreme blur or blackouts—regardless of the higher-level policy.

4.3 L2: Calibration via Contextual Bandits

L1 enforces a rigid safety envelope to prevent failure. Inside this bounded space, the Cerebellum (L2) performs fine-grained optimization guided by feedback from the Cerebrum-F (L3). By correlating sensor actions with downstream task

confidence, L2 acquires an *internal model* that links scene conditions to the ideal exposure strategy.

Formulation. We instantiate L2 as a **Contextual Multi-Armed Bandit (CMAB)**. Unlike full reinforcement learning, which models state transitions, our formulation reflects the physical reality that camera actions (exposure/ISO) do not influence future environmental states (ambient light/motion). Thus, the agent focuses on maximizing the *immediate* reward for the current context.

Context and Action. The *context* s_t corresponds to the discretized state tuple defined in L1:

$$s_t = (s_{motion}, s_{light})$$

The *arms* (actions) are defined as discrete relative offsets applied to the baseline indices proposed by the Brainstem (L1). Instead of exploring the entire parameter space, L2 "nudges" the sensor settings:

$$a_t = (\Delta\text{Idx}_{ISO}, \Delta\text{Idx}_{Exp}) \in \{-1, 0, +1\}^2$$

Crucially, the final applied setting is the result of this nudge clamped by the Brainstem's safety envelope $\mathcal{S}_{safe}$. If L2 suggests an unsafe action (e.g., increasing exposure during fast motion), L1 intercepts and overrides it, ensuring exploration never compromises safety.

Reward Signal. Consistent with the lap-based task objective, we compute the reward r_t upon the completion of each cycle. The reward is a weighted sum of the downstream model's confidence and a signal quality penalty:

$$r_t = \alpha \cdot \max(\text{Conf}_{L3}) + (1 - \alpha) \cdot \text{Score}_{Quality}$$

where Conf_{L3} is the maximum confidence score returned by Cerebrum-F (L3) during the lap, and $\text{Score}_{Quality}$ penalizes motion blur. We set $\alpha = 0.9$ to prioritize recognition certainty.

Policy Consolidation. To mimic biological motor learning, we implement a consolidation mechanism. When the bandit consistently selects a specific optimal action for a given context (i.e., convergence), this context-action pair is cached into a persistent *policy lookup table*. This effectively "engraves" the learned internal model into the reactive layer. Consequently, the L1/L2 can *dynamically switch* to optimal configurations as the physical context (s_{motion}, s_{light}) changes, *decoupling* sensor control from the heavy semantic loop of the Cerebrum.

4.4 L3/L4: Hybrid Cortex and Routing

The cortical planes are instantiated by two distinct classifiers, creating a split architecture that balances real-time responsiveness with recognition capability.

Cerebrum-F (L3). We employ MobileNetV2 [44], pre-trained on ImageNet, as the compact local classifier. Running directly on the mobile device, L3 processes frames in real-time. It acts as the first line of interpretation, outputting a predicted object class and a confidence score for every frame captured during the sensing loop.

Cerebrum-D (L4). For deep, remote inference, we utilize the Gemini API (specifically the `gemini-2.5-flash-lite` model) [9]. To ensure machine-parsable outputs suitable for system integration, we explicitly instruct the model using a structured prompt:

```
"You are an expert image classifier trained
on ImageNet-1k. Identify the main object in
this image. Output strictly JSON: {"class_name":
"...", "confidence(%)": "..."}"
```

While this path incurs network latency, it serves as a high-capability oracle for ambiguous samples that the local model fails to recognize with certainty.

Threshold-Based Aggregation and Routing. ATI enables a routing mechanism tightly integrated with the task structure. In the general ATI architecture defined in §3.4, routing is constrained by strict deadlines. However, for this prototype classification task, we prioritize accuracy over strict timing constraints, thus relaxing the deadline check. Instead, we focus on minimizing the high latency cost of remote inference—empirically, L4 round-trips take approximately 10× longer than local L3 inference.

Rather than evaluating every frame remotely, the system aggregates predictions over a single sensing cycle (one lap). Specifically, it tracks the frame with the maximum L3 confidence score recorded during the cycle. Upon lap completion, the Router evaluates this peak confidence alongside the frame's quality, instantiated here as a Blur Score (variance of the Laplacian):

- **Local Acceptance:** If the peak confidence $\geq 50\%$, the system accepts the corresponding L3 prediction as final.
- **Escalation:** If the peak confidence $< 50\%$ *and* the Blur Score exceeds a validity threshold τ_{valid} (i.e., the image is sufficiently sharp), the frame is escalated to L4.
- **Rejection:** If the Blur Score falls below τ_{valid} (indicating severe motion blur), the frame is *discarded locally*. This prevents wasting costly cloud inference on data that is inherently unrecoverable.

This logic minimizes unnecessary cloud overhead, reserving the slow-but-powerful L4 strictly for cases where optimized local sensing (L1/L2) fails to produce a confident result.

5 Evaluation

We evaluate the prototype to answer three key questions: (1) Can the Cerebellum effectively learn to stabilize inputs? (2) Does active sensing improve downstream accuracy compared to standard auto-exposure? (3) Does ATI reduce reliance on the costly remote cortex?

5.1 Model Configurations

To isolate the impact of each ATI component, we compare six configurations by combining three sensor control strategies (AE, L1, L1/L2) with two inference paths (Local L3, Hybrid L3-L4). This factorial design allows us to systematically disentangle three key factors: (i) the impact of safety reflexes (AE vs. L1), (ii) the value of learned stabilization (L1 vs. L1/L2), and (iii) the efficiency of the split cortex (L3 vs. L3-L4).

Baselines (Standard Android Stack).

- **AE-L3 / AE-L3-L4:** The default smartphone setup using built-in Auto-Exposure (AE). We evaluate it both in a standalone local mode (AE-L3) and a hybrid mode (AE-L3-L4) that offloads uncertain frames to Gemini.

Ablations (Safety Reflex Only).

- **L1-L3 / L1-L3-L4:** Replaces AE with our rule-based Brainstem (L1), but without the learned Cerebellum. This isolates the benefit of hard safety constraints (e.g., anti-blur caps) before any learning takes place.

ATI (Proposed).

- **L1/L2-L3 (Device-Only ATI):** Full sensor control (L1+L2) constrained to local inference. This highlights the Cerebellum's ability to maximize accuracy purely through better sensing, without cloud assistance.
- **L1/L2-L3-L4 (Full ATI):** The complete system. L1 and L2 actively stabilize capture, L3 resolves common frames locally, and the Router escalates ambiguous ones to L4. This unifies active sensing with tiered inference.

5.2 Results: L2 Learning Dynamics

To verify the adaptation capability of the Cerebellum, we conducted a continuous learning experiment. We placed a single object on the track and maintained a sustained dark condition (< 10 lux) to force the system to deviate from standard auto-exposure limits. Figure 4 plots the evolution of the camera control parameters and the resulting confidence scores over 550 consecutive laps.

Exploration to Convergence. Initially (Laps 0–180), the bandit explores within the Brainstem's safety envelope. It often selects long exposures which cause motion blur, resulting in low confidence. A distinct phase shift occurs around Lap 200: the agent discovers that maximizing ISO (>3000) permits minimizing exposure time. This policy effectively freezes motion, causing the rolling average confidence (red line) to climb sharply and stabilize around 70–80%.

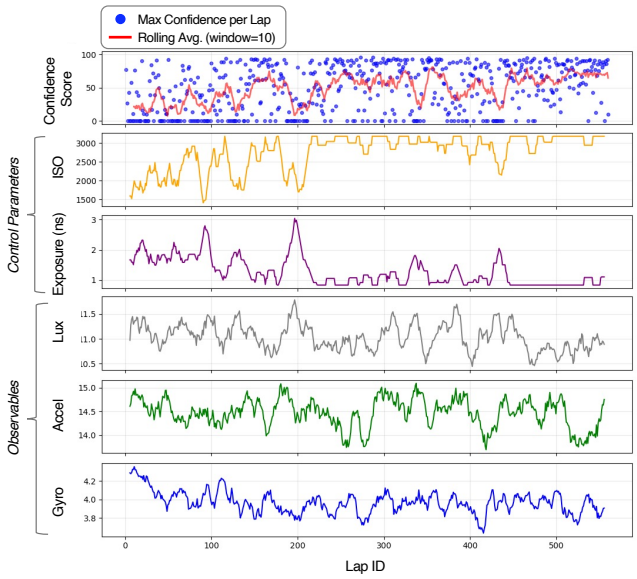

Figure 4: Evolution of camera parameters and confidence scores during Cerebellar learning. Starting from a heuristic baseline (L1), the agent employs Contextual Bandits (L2) to adaptively find the optimal ISO and exposure settings for a dynamic, low-light scene. The rolling average of the confidence score (red line) verifies the system's improved perception capability as it learns the environment over consecutive laps.

Steady State Stability. Post-convergence (Laps 250+), the system consistently adheres to this high-ISO, short-exposure configuration. The remaining fluctuations in confidence are due to geometric factors (e.g., occlusion during turns) rather than exposure instability, confirming that the internal model has successfully stabilized the sensory stream.

Value of Active Intervention. In contrast, repeating this experiment under bright illumination yielded only marginal gains over the baseline (data omitted). This confirms that L2 provides maximum value in the "physical regime" where tight constraints (low light + motion) demand active trade-offs, identifying when active intervention is worth the cost.

5.3 Results: Accuracy and Efficiency

Table 3 summarizes the classification accuracy and L4 offloading rates across eight target objects. To ensure consistent evaluation, we merged semantically equivalent labels (e.g.,

Table 3: Comparison of classification performance and resource usage across different configurations. L4 refers to the remote cortex implemented via a Cloud API. The L4 Call Rate denotes the percentage of frames offloaded to the remote cortex. Note that our proposed method (ATI) significantly improves accuracy while reducing dependency on the expensive L4 layer compared to the baseline.

Method	Model Config	Per-Class Accuracy (%)								Total	L4 Call Rate (%)
		Teddy	Racket	Tennis Ball	Ping-pong Ball	Orange	Carton	Bottle	Laptop		
Baseline	AE-L3	4.0	10.0	74.0	2.0	4.0	0.0	46.0	0.0	17.5	-
	AE-L3-L4	22.0	18.0	78.0	2.0	6.0	12.0	60.0	10.0	26.0	41.8
ATI	L1-L3	36.0	34.0	82.0	0.0	14.0	4.0	84.0	0.0	31.8	-
	L1-L3-L4	50.0	46.0	86.0	0.0	20.0	18.0	88.0	26.0	41.8	28.3
	L1/L2-L3	62.0	48.0	96.0	16.0	28.0	14.0	96.0	0.0	45.0	-
	L1/L2-L3-L4 (Ours)	**68.0**	**68.0**	**96.0**	**18.0**	**32.0**	**18.0**	**98.0**	18.0	**52.0**	**26.3**

mapping 'notebook' to 'laptop' and 'box' to 'carton') prior to calculating accuracy. We compare our proposed ATI system against the standard hybrid baseline (AE-L3-L4), which relies on the phone's built-in auto-exposure.

Accuracy and Efficiency Gains. The proposed method (L1/L2-L3-L4) achieves a total accuracy of 52.0%, significantly outperforming the baseline's 26.0%. Notably, this performance gain is achieved with greater efficiency; ATI reduces the L4 call rate from 41.8% to 26.3%. This confirms that active sensor control improves the quality of raw images, allowing the lightweight local cortex (L3) to classify more frames confidently without needing help from the cloud. These results illustrate how the Brainstem and Cerebellum stabilize sensing to maximize the efficacy of the local cortex, effectively reserving the deep remote cortex for only the strictly necessary, hardest frames.

Adaptation to Object Difficulty. The system shows distinct behaviors depending on object difficulty. For easier objects like *Tennis Ball* and *Water Bottle*, the Cerebellum (L2) successfully learns optimal exposure policies, boosting local accuracy to near perfection (96%–98%) and effectively eliminating L4 usage (0%–4%). In contrast, for challenging objects like *Ping-pong* (small and featureless) or *Laptop* (complex texture), the system maintains a higher L4 call rate (42%–48%). However, even in these difficult cases, the combination of sensor control and selective offloading yields higher final accuracy than the baseline, demonstrating ATI's robustness in handling diverse visual targets.

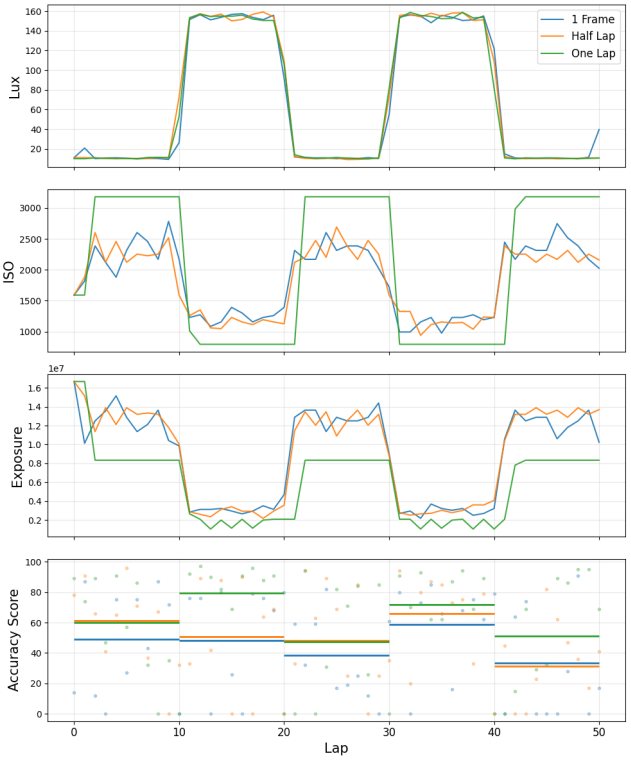

Figure 5: Impact of Observation Window Size on L2 Performance. The global average (One Lap) maximizes perception quality by filtering noise. Even with minimal information (1 Frame), the agent maintains a functional success rate, robustly adapting to dark-bright transitions.

5.4 Robustness to Environmental and Budget Constraints

Experimental Setup: Dynamic Lighting. To evaluate adaptability, we cycled ambient illumination between dark (< 10 Lux) and bright (> 150 Lux) every 10 laps while tracking a 'Teddy Bear' at constant speed. Figure 5 confirms that L2 detects these transitions solely from visual inputs, dynamically adjusting exposure without external triggers to maintain optimal visibility.

Impact of Observation Window Size. To assess stability under latency constraints, we compared three window sizes. Restricting the number of past frames mimics scenarios with tight latency requirements.

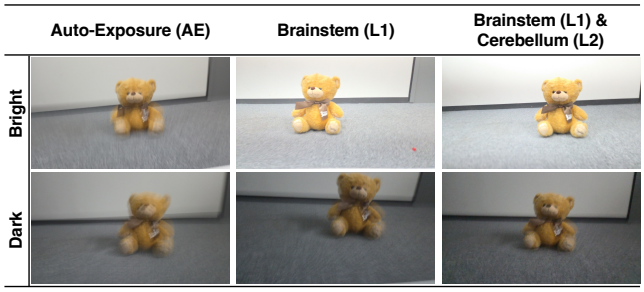

Figure 6: Qualitative comparison of captured frames. (Left) Standard AE fails to account for motion, resulting in severe blur. (Middle) L1 eliminates blur via safety caps but produces under-exposed images in low light due to conservative heuristics. (Right) L1+L2 maintains sharpness while optimizing brightness, demonstrating that the learned policy navigates the safety envelope more effectively than static rules.

Benefit of Global Context (One Lap). Figure 5 shows that the One Lap context yields the best performance (Mean Confidence: 61.5, Success: 68%). This global averaging filters sensor noise, enabling the RL agent to converge on a robust policy that accurately tracks the lighting cycles.

Resilience under Minimal Information. Reducing context lowers stability: mean scores drop to 50.8 (Half Lap) and 45.7 (1 Frame) as the agent overreacts to noise. Notably, even with the minimal 1 Frame input, the agent maintains a 54% success rate, demonstrating the policy's efficacy even without temporal smoothing.

5.5 Qualitative Analysis

Figure 6 visualizes the impact of the active sensing stack. Standard AE (left) fails in dynamic settings, producing severe motion blur regardless of lighting. L1 (middle) successfully eliminates blur by enforcing safety limits; in bright conditions (top row), this reflex alone is sufficient to yield clear images. However, in dark conditions (bottom row), L1's conservative heuristics lead to underexposure. The full L1+L2 stack (right) overcomes this by learning to optimize ISO within the safety envelope, recovering object visibility without re-introducing blur. This confirms that while reflexes guarantee safety, the learned internal model is essential for optimal visibility under challenging constraints.

6 Discussion and Open Challenges

ATI reframes physical AI not as a model-scaling race, but as an architectural contract: protect sensing locally, deliver the first useful decision under strict deadlines, and buy depth only when the cost is justified. While we demonstrated this primarily through vision, the principles of latency lanes and sensor–cortex decoupling extend far beyond cameras. In this section, we discuss how ATI generalizes to other modalities and highlight the missing primitives the systems community must build to make this architecture standard practice.

6.1 Generalization beyond Vision

The core ATI principle—decomposing intelligence into safety reflexes (Brainstem), predictive calibration (Cerebellum), and hierarchical reasoning (Cerebrum)—aligns with the biological organization of sensory processing across all modalities (Table 1). While our prototype instantiated Visual ATI, the same architectural contract applies to audio, touch, and proprioception. High-level perception models (L3/L4) are already well-established for these modalities, such as Transformer architectures for audio and speech processing [15, 41] and graph-based neural models for tactile perception [7, 13]. However, the critical system gap lies in the lower layers: current stacks lack the active mechanisms to stabilize physical signals *before* they reach these heavy models. Therefore, we focus below on defining the missing Brainstem and Cerebellum roles for each modality.

Auditory ATI (The "Smart Cochlea"). Conventional audio pipelines are largely passive at the sensor level. An ATI-based stack would actively reshape signals *before* heavy DSP.

- **Brainstem (L1):** Enforces hardware-level *clipping protection*. Analogous to the stapedius (acoustic) reflex [40] that dampens loud sounds to protect the cochlea, L1 adjusts analog gain on short time scales to prevent sensor saturation during transient spikes.
- **Cerebellum (L2):** Mimics outer hair cells [1] by dynamically tuning *active gain* and steering *beamforming arrays*. In our stack, L2 tunes active gain and steers beamforming arrays in real time. It performs "physical attention," boosting the SNR of a target speaker highlighted by higher auditory centers, rather than amplifying environmental noise blindly.

Tactile ATI (Active Touch). Tactile sensing is inherently interactive: resolving texture and shape typically requires motion and exploratory contact [25, 30].

- **Brainstem (L1):** Monitors contact pressure at high frequency (>1 kHz) to trigger *reflex withdrawal* upon sharp impact, analogous to nociceptive withdrawal reflexes that rapidly remove a limb from harmful stimuli without cortical involvement [47]. This protects the tactile sensor and end-effector before the planner can react.
- **Cerebellum (L2):** Optimizes bandwidth via *event-driven ROI selection* (reading only active taxels) and modulates

contact force (e.g., rubbing harder/softer) to improve signal-to-noise ratio for downstream perception.

Proprioceptive ATI (The "Smart IMU"). Standard IMUs suffer from drift and rely on static filtering. ATI instead enables context-aware state estimation.

- **Brainstem (L1):** Enforces *kinematic limits*. It triggers emergency stops if acceleration or joint velocity exceeds hardware ratings, acting as a "digital spinal cord" that guarantees basic physical safety.
- **Cerebellum (L2):** Dynamically tunes filter parameters (e.g., Kalman gains) and estimates *sensor bias* online. Rather than relying on one-shot calibration, L2 learns to suppress vibration noise during motion while boosting sensitivity when stationary, mirroring the adaptability of the biological vestibular system.

6.2 A Systems Agenda for Physical AI

To transition ATI from a design pattern into a portable ecosystem, the mobile systems community must look beyond individual models and standardize the architectural substrate. We identify three missing primitives.

Standardizing Sensor-Cortex Contracts. Architectures scale only when interfaces stabilize. We propose three high-leverage standards: (i) **Universal Quality Vectors (QV)**, a schema (e.g., `{blur, saturation, SNR}`) allowing L2 modules to communicate signal health to L3 models across heterogeneous devices; (ii) **Advisory Escalation Protocols**, defining RPCs with strict time-to-live (TTL) semantics to ensure cloud latencies never block local loops; and (iii) **Verifiable Safety Envelopes**, a machine-checkable format for L1 constraints that allows developers to formally verify that a learned policy cannot violate hardware limits.

Metrics Beyond Static Accuracy. Standard CV metrics such as classification accuracy and mAP fail to capture the dynamics of embodied systems. We advocate for reporting: (i) **Time-to-First-Decision (TTFD)**, the P99 latency to the first safe action, penalizing systems that wait for heavy models when a reflex would suffice; (ii) **Prevention Rate**, counting how many irrecoverable sensory faults (e.g., washed-out or saturated frames) were preemptively fixed by L1/L2 at the sensor level; and (iii) **Energy-per-Legible-Bit**, a metric incentivizing efficient physical capture of usable signals over brute-force digital restoration.

Tooling for Reproducibility. Evaluating active sensing is notoriously difficult without physical robots. The community needs *sweep-set benchmarks* tailored to specific modalities. For Visual ATI, this requires datasets of short clips captured under systematic sweeps of exposure, ISO, and focus, coupled with a trace-driven simulator. Establishing such shared infrastructure would allow researchers to train and compare L2 policies offline without requiring identical physical hardware, significantly lowering the barrier to entry.

6.3 Risks and Limitations

Adversarial Sensor Control. Granting learned agents control over hardware exposes new vectors for *sensor denial-of-service*. Adversarial physical patterns (e.g., stroboscopic lights) could mislead the Cerebellum into blinding the sensor or inducing oscillation. This necessitates treating the Brainstem (L1) as a *Trusted Computing Base (TCB)*, enforcing immutable safety limits (e.g., exposure caps) that no learned policy can override, regardless of the input.

Scope of Applicability. ATI is not a universal replacement for monolithic models. It yields gains primarily when the sensor exposes *controllable* trade-offs and the task exhibits a *skewed difficulty distribution* (i.e., most frames are easy and only a small fraction are hard). In static environments with ample power and compute, a single monolithic model may be preferable. ATI specifically targets the "physical regime" where resource constraints are tight, dynamics are fast, and the cost of sensing is comparable to the cost of inference.

7 Conclusion

AI's transition from the web to the physical world runs into hard boundaries: sensing errors are often irrecoverable, and decisions must satisfy strict deadlines. We argue that reliable physical AI is not merely a model-scaling problem but an architectural one. ATI proposes a bio-inspired contract that formalizes this reliability: a Brainstem and Cerebellum that actively protect and stabilize the signal at the sensor edge, coupled with a split Cerebrum that provides fast local decisions while routing only the hardest queries to the cloud. By treating placement as policy and enforcing strict latency lanes, ATI aims to keep systems safe by reflex and refined by escalation.

This architecture aims to turn "AI at the edge" from a slogan into a programmable discipline. Instead of training opaque monolithic models, ATI factorizes learning by configuring and auditing safety reflexes, learning calibration from feedback, and distilling cloud wisdom into compact local heads. We invite the systems community to standardize the missing primitives identified in this work, namely universal quality vectors, advisory escalation APIs, and verifiable safety envelopes. Ultimately, the path to robust embodied intelligence lies in a simple yet rigorous principle: protect the signal first, stabilize it next, interpret it quickly, and escalate only when it pays.

References

[1] Jonathan Ashmore. 2008. Cochlear outer hair cell motility. *Physiological reviews* 88, 1 (2008), 173–210.

[2] Eunsu Baek, Sung-hwan Han, Taesik Gong, and Hyung-Sin Kim. 2025. Adaptive Camera Sensor for Vision Models. In *The Thirteenth International Conference on Learning Representations*.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Herman Bruyninckx. 2001. Open robot control software: the OROCOS project. In *Proceedings 2001 ICRA. IEEE international conference on robotics and automation (Cat. No. 01CH37164)*, Vol. 3. IEEE, 2523–2528.

[5] Kaiyuan Chen, Kush Hari, Trinity Chung, Michael Wang, Nan Tian, Christian Juette, Jeffrey Ichnowski, Liu Ren, John Kubiatowicz, Ion Stoica, et al. 2024. FogROS2-FT: Fault Tolerant Cloud Robotics. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1390–1397.

[6] Kaiyuan Chen, Nan Tian, Christian Juette, Tianshuang Qiu, Liu Ren, John Kubiatowicz, and Ken Goldberg. 2025. Fogros2-plr: Probabilistic latency-reliability for cloud robotics. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 16290–16297.

[7] Lun Chen, Yingzhao Zhu, and Man Li. 2024. Tactile-GAT: tactile graph attention networks for robot tactile perception classification. *Scientific Reports* 14, 1 (2024), 27543.

[8] Claudia Clopath, Aleksandra Badura, Chris I De Zeeuw, and Nicolas Brunel. 2014. A cerebellar learning model of vestibulo-ocular reflex adaptation in wild-type and mutant mice. *Journal of Neuroscience* 34, 21 (2014), 7203–7215.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[10] Eduardo Cuervo, Aruna Balasubramanian, Dae-ki Cho, Alec Wolman, Stefan Saroiu, Ranveer Chandra, and Paramvir Bahl. 2010. Maui: making smartphones last longer with code offload. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*. 49–62.

[11] Michael Tri H Do. 2019. Melanopsin and the intrinsically photosensitive retinal ganglion cells: biophysics to behavior. *Neuron* 104, 2 (2019), 205–226.

[12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *International Conference on Machine Learning*. PMLR, 8469–8488.

[13] Wen Fan, Hongbo Bo, Yijiong Lin, Yifan Xing, Weiru Liu, Nathan Lepora, and Dandan Zhang. 2022. Graph neural networks for interpretable tactile sensing. In *2022 27th International Conference on Automation and Computing (ICAC)*. IEEE, 1–6.

[14] Biyi Fang, Xiao Zeng, and Mi Zhang. 2018. Nestdnn: Resource-aware multi-tenant on-device deep learning for continuous mobile vision. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 115–127.

[15] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Interspeech 2021*. 571–575. doi:10.21437/Interspeech.2021-698

[16] Lixiang Han, Zimu Zhou, and Zhenjiang Li. 2024. Pantheon: Preemptible multi-dnn inference on mobile edge gpus. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 465–478.

[17] Seungyeop Han, Haichen Shen, Matthai Philipose, Sharad Agarwal, Alec Wolman, and Arvind Krishnamurthy. 2016. Mcdnn: An approximation-based execution framework for deep stream processing under resource constraints. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. 123–136.

[18] J Johanna Hopp and Albert F Fuchs. 2004. The characteristics and neuronal substrate of saccadic eye movement plasticity. *Progress in neurobiology* 72, 1 (2004), 27–53.

[19] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Hao-Shu Fang, et al. 2023. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782* (2023).

[20] Jeffrey Ichnowski, Kaiyuan Chen, Karthik Dharmarajan, Simeon Adebola, Michael Danielczuk, Víctor Mayoral-Vilches, Nikhil Jha, Hugo Zhan, Edith Llontop, Derek Xu, et al. 2023. FogROS2: An Adaptive Platform for Cloud and Fog Robotics Using ROS 2. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5493–5500.

[21] Masao Ito. 1998. Cerebellar learning in the vestibulo–ocular reflex. *Trends in cognitive sciences* 2, 9 (1998), 313–321.

[22] Masao Ito. 2008. Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience* 9, 4 (2008), 304–313.

[23] Joo Seong Jeong, Jingyu Lee, Donghyun Kim, Changmin Jeon, Changjin Jeong, Youngki Lee, and Byung-Gon Chun. 2022. Band: coordinated multi-dnn inference on heterogeneous mobile processors. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 235–247.

[24] Fucheng Jia, Shiqi Jiang, Ting Cao, Wei Cui, Tianrui Xia, Xu Cao, Yuanchun Li, Qipeng Wang, Deyu Zhang, Ju Ren, et al. 2024. Empowering in-browser deep learning inference on edge through just-in-time kernel optimization. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 438–450.

[25] Roland S Johansson and J Randall Flanagan. 2009. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience* 10, 5 (2009), 345–359.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[27] Victor AF Lamme and Pieter R Roelfsema. 2000. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences* 23, 11 (2000), 571–579.

[28] Michael Land and Benjamin Tatler. 2009. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[30] Susan J Lederman and Roberta L Klatzky. 2009. Haptic perception: A tutorial. *Attention, Perception, & Psychophysics* 71, 7 (2009), 1439–1459.

[31] Kyunghyun Lee, Ukcheol Shin, and Byeong-Uk Lee. 2024. Learning to control camera exposure via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2975–2983.

[32] Stephen G Lisberger. 2015. Visual guidance of smooth pursuit eye movements. *Annual review of vision science* 1, 1 (2015), 447–468.

[33] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics* (2025).

[34] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. 165–178.

[35] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093* (2024).

[36] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. 2022. Robot operating system 2: Design, architecture, and uses in the wild. *Science robotics* 7, 66 (2022), eabm6074.

[37] A David Milner and Melvyn A Goodale. 2008. Two visual systems re-viewed. *Neuropsychologia* 46, 3 (2008), 774–785.

[38] David Milner and Mel Goodale. 2006. *The visual brain in action*. Vol. 27. Oup Oxford.

[39] Emiliano Miluzzo, Nicholas D Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B Eisenman, Xiao Zheng, and Andrew T Campbell. 2008. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*. 337–350.

[40] Aage R Møller. 2012. *Hearing: anatomy, physiology, and disorders of the auditory system*. Plural Publishing.

[41] Ngoc-Quan Pham, Thai Son Nguyen, Jan Niehues, Markus Müller, and Alexander H. Waibel. 2019. Very Deep Self-Attention Networks for End-to-End Speech Recognition. In *Proc. Interspeech*.

[42] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, 5.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[45] Mahadev Satyanarayanan, Paramvir Bahl, Ramón Caceres, and Nigel Davies. 2009. The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing* 8, 4 (2009), 14–23.

[46] Anna Sedda and Federica Scarpina. 2012. Dorsal and ventral streams across sensory modalities. *Neuroscience bulletin* 28, 3 (2012), 291–300.

[47] V Skljarevski and NM Ramadan. 2002. The nociceptive flexion reflex in humans–review article. *Pain* 96, 1-2 (2002), 3–8.

[48] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. 2025. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020* (2025).

[49] Simon Thorpe, Denis Fize, and Catherine Marlot. 1996. Speed of processing in the human visual system. *nature* 381, 6582 (1996), 520–522.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[51] John Von Neumann. 1993. First Draft of a Report on the EDVAC. *IEEE Annals of the History of Computing* 15, 4 (1993), 27–75.

[52] Zhonglin Yang, Hao Fang, Huanyu Liu, Junbao Li, Yutong Jiang, and Mengqi Zhu. 2024. Active Visual Perception Enhancement Method Based on Deep Reinforcement Learning. *Electronics* 13, 9 (2024), 1654.

[53] Alfred L Yarbus. 2013. *Eye movements and vision*. Springer.

[54] Juheon Yi, Sunghyun Choi, and Youngki Lee. 2020. EagleEye: Wearable camera-based person identification in crowded urban spaces. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[55] Juheon Yi and Youngki Lee. 2020. Heimdall: mobile GPU coordination platform for augmented reality applications. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[56] Andrew J Zele, Beatrix Feigl, Simon S Smith, and Emma L Markwell. 2011. The circadian response of intrinsically photosensitive retinal ganglion cells. *PLOS one* 6, 3 (2011), e17860.

[57] Shuyang Zhang, Jinhao He, Yilong Zhu, Jin Wu, and Jie Yuan. 2024. Efficient Camera Exposure Control for Visual Odometry via Deep Reinforcement Learning. *IEEE Robotics and Automation Letters* (2024).

[58] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. 2019. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc. IEEE* 107, 8 (2019), 1738–1762.

[59] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*. PMLR, 2165–2183.