AI Should Sense Better, Not Just Scale Bigger: Adaptive Sensing as a Paradigm Shift

Eunsu Baek[1], Keondo Park[1], JeongGil Ko[2], Min-hwan Oh[1], Taesik Gong[3], and Hyung-Sin Kim[*1]

[1]Graduate School of Data Science, Seoul National University, Seoul, South Korea
[2]School of Integrated Technology, Yonsei University, Seoul, South Korea
[3]Department of Computer Science and Engineering, UNIST, Ulsan, South Korea
[1]{beshu9407, gundo0102, minoh, hyungkim}@snu.ac.kr
[2]jeonggil.ko@yonsei.ac.kr
[3]taesik.gong@unist.ac.kr

Abstract

Current AI advances largely rely on scaling neural models and expanding training datasets to achieve generalization and robustness. Despite notable successes, this paradigm incurs significant environmental, economic, and ethical costs, limiting sustainability and equitable access. Inspired by biological sensory systems, where adaptation occurs dynamically at the input (e.g., adjusting pupil size, refocusing vision)—we advocate for *adaptive sensing* as a necessary and foundational shift. Adaptive sensing proactively modulates sensor parameters (e.g., exposure, sensitivity, multimodal configurations) at the input level, significantly mitigating covariate shifts and improving efficiency. Empirical evidence from recent studies demonstrates that adaptive sensing enables small models (e.g., EfficientNet-B0) to surpass substantially larger models (e.g., OpenCLIP-H) trained with significantly more data and compute. We (i) outline a roadmap for broadly integrating adaptive sensing into real-world applications spanning humanoid, healthcare, autonomous systems, agriculture, and environmental monitoring, (ii) critically assess technical and ethical integration challenges, and (iii) propose targeted research directions, such as standardized benchmarks, real-time adaptive algorithms, multimodal integration, and privacy-preserving methods. Collectively, these efforts aim to transition the AI community toward sustainable, robust, and equitable artificial intelligence systems.

1 Introduction

Early neural-network research drew inspiration from cortical neurons, implicitly framing intelligence as a predominantly brain-centered phenomenon [44, 58]. Correspondingly, recent advancements in artificial intelligence (AI) have predominantly relied on scaling model size and expanding training datasets [69, 2, 73, 8]. Although effective, this *model-centric paradigm* introduces significant and unsustainable challenges, such as massive computational demands that exacerbate environmental degradation [12, 53], socioeconomic inequities due to resource concentration in a few well-funded institutions [10, 11], and persistent failures in generalizing under real-world domain shifts [25, 35].

In contrast, biological cognition is fundamentally embodied: the brain operates within an integrated system comprising sensory organs (e.g., vision, hearing, and touch) and musculature, which forms critical perception-action loops essential for survival and evolutionary success. For robust perception, human sensory systems dynamically adapt at the sensor level both before and during cortical processing. For instance, humans correct optical blur by wearing corrective lenses rather than retraining their neural circuits with thousands of out-of-focus images. Similarly, when facing low contrast

[*] Corresponding author

or glare, humans instinctively squint, rapidly narrowing the pupil to sharpen depth-of-field and suppress stray light. Thus, intelligence emerges from the coordinated evolution and dynamic interplay between sensory organs, neural pathways, and adaptive behaviors, not solely from neural complexity alone [33, 65]. This biological principle highlights that effective perception and generalization in AI depend **not only on neural capacity but also on dynamic, real-time sensor adaptation** throughout the entire perception pipeline.

However, current AI methodologies still primarily address distribution shift by *scaling* neural architectures and datasets [49, 15, 2, 57], leaving the sensing interface unchanged. A few existing sensor-aware strategies, such as active perception, which repositions sensors or robots to optimize viewpoints [9, 64], and sensor-fusion budgeting, which schedules when and what modalities to activate [70], operate at the system or motion-planning level. They neglect how raw analog signals are digitized in the first place to best serve the downstream model. Only recently have benchmarks begun isolating the impact of intrinsic sensor settings, such as exposure and gain, on recognition accuracy [7, 6]. Inspired by these insights, the concept of **test-time input adaptation** has emerged: dynamically optimizing sensor parameters frame-by-frame to deliver model-friendly data. A first prototype, Lens [6], significantly outperforms traditional (human-oriented) auto-exposure methods, maintaining high classification accuracy despite a 50× reduction in model size. Preliminary evidence suggests even greater potential: with ideal sensor adaptation, a lightweight 5M-parameter EfficientNet-B0 [42] can surpass the 632M-parameter OpenCLIP-H [29] trained on 160× more training data [7]. These early results highlight both the promise of adaptive sensing and the urgency for systematic exploration.

While initial empirical demonstrations come from image-classification tasks, adaptive sensing is broadly applicable across diverse domains. By shaping incoming photons, acoustic waves, or tactile forces *at the sensor*, it boosts efficiency and robustness in fields such as medical diagnostics [54, 4], autonomous driving [60], surveillance [77], and environmental monitoring [3, 43]. This capability is becoming critical as AI increasingly transitions from artificial, controlled simulations to **physical, embodied deployments** [51]. In real-world robots and wearable devices, identical models may receive data from lenses, CCDs, MEMS microphones, or tactile skins, each differing across vendor, firmware, and production batch. These subtle differences introduce domain shifts, either explicit or latent, that are difficult to address through model advancement alone.

Recent embodied-AI benchmarks in locomotion [62, 39], household manipulation [66], and interactive task execution [81] exemplify an urgent reality, where the future AI must sense, reason, and interact from and with the real world within strict real-time and on-device constraints. Physical AI platforms from micro-drones to neural prostheses face inherent limits in size, energy, and thermal dissipation, suggesting limits to the traditional “bigger model, bigger dataset” paradigm. Adaptive sensing provides a complementary solution: by reshaping inputs directly at the source, it delivers model-friendly signals that enable smaller networks to thrive. Just as the human visual system couples rapid eye movements to visual cognition, next-generation AI systems must integrate closed-loop, real-time sensor control and preprocessing directly into their inference pipeline.

Our Position. Inspired by the principle in biological sensory systems, we argue that **AI research must transition away from an exclusively model-centric paradigm toward prioritizing adaptive, input-level optimization as a first-class concern.** Adaptive sensing is more than an incremental advance. Rather, it represents a critical, paradigm-level shift toward sustainable, equitable, and robust AI. Sections 2-7 articulate key research directions, evaluation frameworks, and interdisciplinary opportunities essential to establish adaptive sensing as a foundational pillar of future AI.

2 Current State of AI: Limitations of the Model-Centric Paradigm

The dominant approach in modern AI emphasizes scaling models and expanding datasets to improve robustness and generalization [12, 11]. Although this model-centric paradigm has yielded notable performance gains, it faces fundamental limitations that increasingly threaten its long-term viability:

- **Environmental and Computational Cost:** Modern AI models, particularly large-scale models such as GPT-3/4 [12, 2], and advanced multimodal models [69, 72, 73], require massive computational resources for training [16]. Training GPT-3 alone requires approximately 1.287 GWh of electricity, emitting roughly 552 tons of CO_2, comparable to driving a passenger vehicle for more than one million kilometers [53, 67]. The ongoing trend toward even larger models [34, 26] exponentially escalates these environmental costs, posing significant sustainability challenges.

- **Accessibility and Inequity:** The heavy computational requirements of training and deploying state-of-the-art models (e.g., GPT variants, CLIP, and advanced multimodal transformers) restrict access primarily to organizations with substantial financial and computational resources [11]. This centralization widens the global digital divide by disproportionately benefiting wealthy institutions and regions while restricting innovation and participation from smaller-scale researchers, startups, and economically disadvantaged groups. Consequently, the potential diversity of ideas and equitable distribution of AI benefits remain severely constrained.
- **Real-World Generalization Failures:** AI models trained on massive yet static datasets frequently fail to generalize effectively when encountering novel or dynamically changing real-world conditions. Domain shifts–variations in environmental, sensor-specific, and task-specific contexts–are notoriously challenging for traditional data-centric approaches [56]. Existing robustness benchmarks inadequately capture the true complexity of real-world conditions [7]. As a result, models trained under these benchmarks often exhibit significant performance drops when deployed in realistic settings, limiting their reliability in critical applications such as autonomous driving [60], medical diagnostics [54, 4] and environmental monitoring [3, 43].
- **Economic Sustainability and Scalability:** The continuous increase in model complexity and training data volumes imposes a substantial financial burden [67, 14]. The financial costs associated with advanced computational infrastructure, specialized hardware, and extensive data acquisition quickly become prohibitive. This economic barrier hampers the scalability and widespread adoption of advanced AI technologies, especially in resource-constrained settings or smaller-scale enterprises.
- **Ethical and Societal Concerns:** Large-scale models inherently risk amplifying biases present in extensive, complex training datasets [24, 36]. As data volumes grow, auditing, identifying, and mitigating embedded biases becomes increasingly difficult [61]. Without rigorous oversight, these biases can propagate societal harm, perpetuating stereotypes, inequities, and systemic prejudices [10]. Addressing these ethical implications through a model-centric lens alone remains insufficient, reinforcing the need for alternative paradigms.

Collectively, these limitations emphasize the urgency of moving beyond a solely model-centric paradigm toward strategies that integrate smarter, adaptive sensing and context-aware optimization.

3 Adaptive Sensing as a Necessary Paradigm Shift

Overcoming the limitations of the model-centric paradigm requires treating sensor-level adaptability as a first-class design principle. While biological sensors embed rapid, energy-proportional adaptation in hardware, modulating gain, bandwidth, and spatial resolution before neural inference begins, modern artificial sensors, such as cameras, microphones, and haptic arrays, **remain predominantly static** (Table 1). Bridging this gap through adaptive sensing at the input level would enable smaller, faster, and fairer AI models, providing a compelling path toward sustainable and equitable artificial intelligence.

Table 1: Comparison between adaptive human sensors and static artificial sensors.

Modality	Human Sensor (Adaptive)	Artificial Sensor (Static)
Vision	Pupil diameter 2–8 mm ($\sim$16× light gain) in < 200 ms. Dark adaptation restores sensitivity; ciliary muscles refocus 10 cm to infinity. Saccades (3–5 ms) redirect fovea before cortex [42, 78].	Fixed or two-step aperture; ISO/shutter in coarse steps [23]. Autofocus 50–300 ms; quantum efficiency and CFA static. Saturation corrected only post-capture.
Hearing	Active gain control (> 120 dB dynamic range) via outer-hair cells [38]. Real-time efferent feedback loop protects hearing from damage [47]. 3-D localization from binaural delay down to 10 μs.	60-90 dB dynamic range (MEMS mic) with fixed AGC [63]. Limited impulse protection. 3D localization requires multi-element arrays and DSP.
Touch	Sensitivity of 0.3-500+ kPa via mosaic of mechanoreceptors. Spatial acuity up to 0.5mm; temporal resolution $\sim$ 1kHz [74, 32]. Soft, compliant skin spreads out pressure; nociceptors trigger pain reflexes.	Sensitivity of 1-100 kPa via capacitive or piezoresistive arrays. Taxel pitch 1mm; sampling 100-500Hz [22]. Mechanically rigid surface with no pain feedback [30].

3.1 Early Evidence from Image Classification

Recent work on ImageNet-ES [7] and ImageNet-ES-Diverse [6] evaluates real-world covariate shifts induced by controlled sensor variation. Lens [6] (Figure 1) is the first model-friendly, test-time input adaptation framework to mitigate covariate shifts. It operates in a post-hoc, adaptive, and camera-agnostic sensor control manner, dynamically responding to scene characteristics based on VisiT scores to provide optimal image quality for neural networks. The key empirical findings are as follows, which underscore both the promise and the necessity of adaptive sensing methodologies:

- **Accuracy gain:** Adaptive sensing significantly improves accuracy up to 47.58%p without model modification or maintain accuracy despite 50× model size difference.

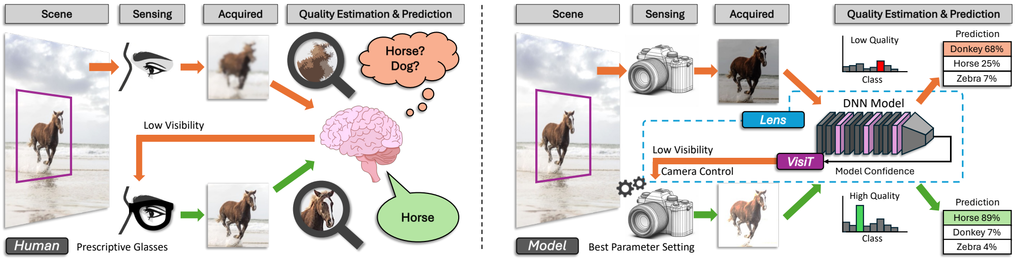


Figure 1: A representative adaptive sensing framework (Lens [6]).

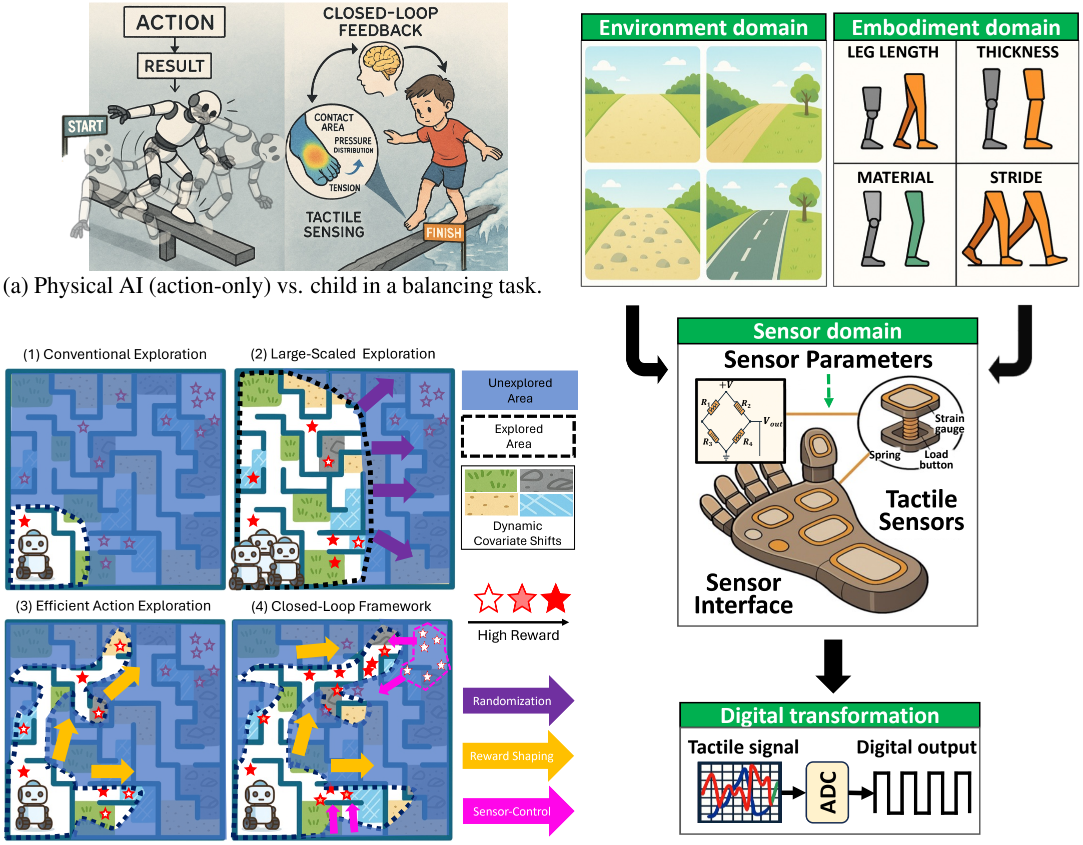
- **Synergy:** Adaptive sensing synergistically integrates with model improvement techniques.
- **Specificity:** Adaptive sensing must be tailored in a model- and scene-specific manner.
- **Human vs. Model optics:** High-quality images for model perception differ from those optimized for human perception.

3.2 Advantages for Real-World Agents and Deployment

Integrating adaptive sensing yields practical advantages in realistic operational settings:

- **Learning perspective:** Unlike simulation environments [45, 21, 71], real-world environments pose challenges due to sensor heterogeneity (cameras, microphones, haptic arrays) and unpredictable conditions (lighting, weather) [40, 13, 31, 28]. Adaptive sensing enables agents to dynamically optimize sensor parameters (e.g., exposure, viewpoint), effectively reducing perceptual uncertainty. This targeted data acquisition approach enhances sample-efficient learning and robust generalization, especially under sparse reward conditions.
- **Engineering and Economic Viability:** Adaptive sensing reduces computational requirements compared to extensive model retraining, well-suited for resource-constrained and embedded systems. The improved data quality at the sensor level reduces infrastructure costs, enabling economically sustainable AI deployment at scale [76].
- **Ethical and Societal Impact:** By enabling targeted, context-aware data collection, adaptive sensing mitigates biases prevalent in large, static datasets [59, 48, 19]. This ensures fairer outcomes in sensitive applications, enhancing public trust and ethical integrity in AI systems.
- **Interdisciplinary Innovation.** Adaptive sensing naturally encourages collaboration among computer scientists, engineers, ethicists, neuroscientists, and policymakers, fostering holistic solutions and comprehensive technological advancements.
- **Broad Domain Applicability:** Adaptive sensing has the potential to practically impact diverse domains that utilize sensor data. Concrete scenarios are as follows:
  - **Humanoid Robotics:** Adaptive multimodal sensor adjustments—such as dynamically modulating visual, auditory, and tactile sensors—to improve real-time balance control, manipulation tasks, and interaction quality in complex, unstructured environments.
  - **Healthcare Diagnostics:** Dynamic adjustment of medical imaging parameters to optimize image quality based on patient-specific characteristics and clinical contexts (e.g., adjusting MRI sequences dynamically to enhance diagnostic accuracy).
  - **Autonomous Vehicles:** Real-time optimization of camera and lidar parameters to rapidly adapt to changing lighting conditions and weather (e.g., automatic adjustments during fog, rain, or sudden brightness changes to enhance road safety).
  - **Agriculture Monitoring:** Adaptive drone sensing systems dynamically adjusting sensor settings to capture high-quality data under varying conditions, such as different crop types, growth stages, or environmental stress factors (e.g., real-time optimization for precise crop health monitoring).
  - **Environmental Monitoring:** Adaptive sensing to optimize sensor settings dynamically for capturing accurate data on air and water quality under varying environmental conditions, thereby improving monitoring accuracy and predictive modeling.

In summary, adaptive sensing represents a necessary paradigm shift for smaller, greener, and fairer AI systems—turning "sense better" into "learn less."



(a) Physical AI (action-only) vs. child in a balancing task.

(b) Exploration in dynamic, sparse-reward settings. (1–3): Motor-only learning vs. (4): Perception-aware learning.

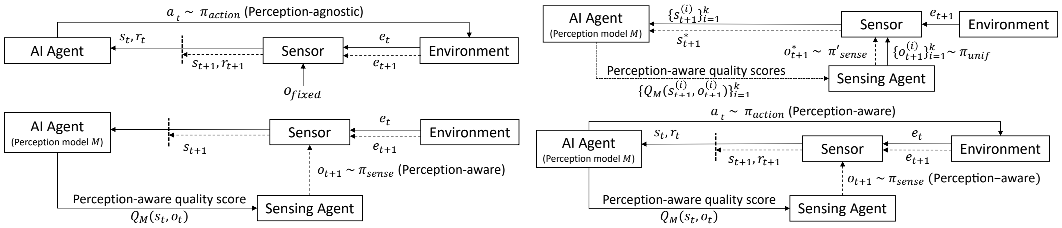
(c) Illustration of covariate shift in embodied AI settings for a balancing task.

Figure 2: Why Closed-Loop Adaptive Sensing Framework is needed for Embodied AI Agents?

4 Adaptive Sensing for Embodied AI: Obstacles and Outlook

This section explores the deeper potential of adaptive sensing within embodied AI, expanding beyond the initial evaluation in static image classification tasks. Current embodied AI systems predominantly rely on *action-centric training* without explicitly considering adaptive sensing, resulting in inefficient learning and limited adaptability [18]. This limitation becomes particularly evident when comparing action-only embodied agents to humans, who seamlessly integrate sensory adaptation with motor actions in a closed-loop manner. For instance, consider a balancing task (Fig. 2a), which involves simultaneous covariate shifts [5, 80, 7] stemming from variations in environmental conditions, sensor interfaces and parameters, and agent morphologies (Fig. 2c), and extremely sparse reward signals relative to the complexity of the multi-sensor, multi-modal exploration space (Fig. 2b). Humans quickly achieve robust performance through efficient closed-loop sensory feedback, requiring only a few training trials and effortlessly adapting to new conditions. In contrast, embodied agents relying solely on motor actions significantly lag in both learning speed and real-world robustness.

Current Gaps in Adaptive Sensing Approaches: Existing adaptive sensing methods, notably Lens [6], predominantly target single-shot perception tasks (Fig. 3b), neglecting scenarios involving continuous interaction and sequential decision-making. These approaches fail to consider the crucial role of ongoing closed-loop feedback between sensing, perception, and action, which is essential for embodied agents operating within dynamic, real-world conditions. Moreover, current methods do not actively utilize model perception feedback to guide sensor parameter exploration. This limitation is especially critical in continuous, sparse-reward scenarios, where effective navigation of the sensor configuration space through closed-loop adaptive sensing is critical for robust and efficient learning. Addressing these gaps is imperative for successfully integrating adaptive sensing with continuous, action-driven learning paradigms in embodied AI.


$a_t \sim \pi_{action}$ (Perception-agnostic)
AI Agent
s_t, r_t
s_{t+1}, r_{t+1}
Sensor
e_t
e_{t+1}
Environment
o_{fixed}
AI Agent
(Perception model M)
s_{t+1}
Sensor
e_t
e_{t+1}
Environment
$o_{t+1} \sim \pi_{sense}$ (Perception-aware)
Perception-aware quality score
$Q_M(s_t, o_t)$
Sensing Agent
AI Agent
(Perception model M)
$\{s_{t+1}^{(i)}\}_{i=1}^{k}$
s_{t+1}^{*}
Sensor
e_{t+1}
Environment
$o_{t+1}^{*} \sim \pi'_{sense}$
$\{o_{t+1}^{(i)}\}_{i=1}^{k} \sim \pi_{unif}$
Perception-aware quality scores
$\{Q_M(s_{t+1}^{(i)}, o_{t+1}^{(i)})\}_{i=1}^{k}$
Sensing Agent
$a_t \sim \pi_{action}$ (Perception-aware)
AI Agent
(Perception model M)
s_t, r_t
s_{t+1}, r_{t+1}
Sensor
e_t
e_{t+1}
Environment
$o_{t+1} \sim \pi_{sense}$ (Perception–aware)
Perception-aware quality score
$Q_M(s_t, o_t)$
Sensing Agent

(c) Perception-Only.

(d) Sensorimotor.

Figure 3: Towards Closed-Loop Adaptive Sensing Framework for Embodied AI Agents.

5 Closed-Loop Framework for Embodied AI Agents: Towards Humanoid

To fill the gaps and integrate adaptive sensing into embodied reinforcement learning (RL), we propose a principled **Adaptive Sensing Framework**, starting from the most fundamental RL setting, a standard MDP formulation [55], and progressively extending it to accommodate single-shot perception, continuous perception tasks, sensorimotor interactions, and ultimately multimodal embodied scenarios, all within a fully closed-loop adaptive structure. The ultimate goal is to establish a robust, adaptive AI pipeline capable of efficiently learning and reliably adapting to dynamic environments.

5.1 RL Agents with Sensing–Environment Interaction

To pave the way for closed-loop adaptive sensing in embodied AI, we first formalize two baseline setups—each reflecting that the agent's state is observed by sensor measurements (via measurement function f) from the interaction between sensor configuration and the environment:

5.1.1 Common Notations

Basic Environment Components. Given state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability P, and reward function R, the environment response distribution $P_E(e_{t+1} \mid s_t, a_t)$ defines the probability of the next environmental state e_{t+1} conditioned on the current state-action pair (s_t, a_t), encompassing all potential covariate shifts arising within the environment domain as a consequence of the agent's actions. Similarly, the reward is generally sampled as: $r_{t+1} \sim R(s_t, \cdot)$ and optionally may incorporate additional contextual information such as sensing or action history.

Sensor Parameter Options Space $\mathcal{O}$ ($\subseteq \mathbb{R}^p$). The set of all possible sensing settings for p parameters. An option $o_t \in \mathcal{O}$ represents settings of sensing parameters at time t. The option $o_{\text{fixed}} \in \mathcal{O}$ denotes a constant parameter configuration, typically used in non-adaptive (fixed) sensing settings.

Sensor Measurement $f(e_t, o_t)$. A sensor operation converting analog signals from the environment e_t into a digital observation s_t under sensor configuration o_t at timestep t. This process inherently captures various covariate shifts resulting from sensor-environment interactions.

Perception-Aware Quality Estimation $Q_M(s_t, \cdot)$. A perception-based metric evaluating how easily an observation s_t can be interpreted by a given perception model (or module) M. For example, if M is an image classification model, the quality metric can be defined as the maximum confidence score (Lens [6]): $Q_M(s_t) := \max(\text{softmax}(M(s_t)))$. Optionally, additional contextual information, such as sensing or action history, can be incorporated into this metric.

Action Policy $\pi_{\textbf{action}}(a_t \mid s_t, \cdot)$. General policy selecting the agent's action based on relevant states and contextual information such as action history (e.g., $\pi_{\text{action}}(a_t \mid s_t, a_{t-1})$).

Sensing Policy $\pi_{\textbf{sense}}(o_{t+1} \mid s_t, \cdot)$. General policy selecting the next sensing configuration conditioned on prior observations and contextual information such as sensing history, and perception-aware quality metrics (e.g., $\pi_{\text{sense}}(o_{t+1} \mid s_t, o_t, Q_M)$).

Conventional RL (MDP) without Adaptive Sensing. (Fig. 3a) We represent the environment as a standard Markov Decision Process (MDP) [55], $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P_E, R)$ assuming non-adaptive (i.e., fixed) sensor configurations. At each timestep t, the agent executes:

1. **State Observation:** $s_t \in \mathcal{S}$.
2. **Action Selection:** $a_t \sim \pi_{action}(a_t \mid s_t)$.
3. **Environment Response and Sensor Measurement (No Sensing Agent):**

$$e_{t+1} \sim P_E(e_{t+1} \mid s_t, a_t), \quad s_{t+1} \sim f(e_{t+1}, o_{fixed}) \tag{1}$$

4. **Reward Collection:** $r_{t+1} \sim R(s_t, a_t)$

Single-Shot Adaptive Sensing (Perception-Only, No Actions, (Fig. 3b)). We define a newly augmented stochastic process $\mathcal{P} := (\mathcal{S}, \mathcal{O}, P'_E, Q_M)$ defined by P'_E which is the environment transition probability similar to P_E defined in Eq.(1), but agnostic to actions. This new stochastic process $\mathcal{P}$ captures ways in which sensing configurations are adaptively selected per-timestep for single-shot perception tasks (e.g., image classification tasks).

Concretely, at timestep t, the agent performs:

1. **State Candidate Observation:** For a given candidate sensor configuration o_{t+1} at timestep $t+1$ and the current state s_t, a candidate sensor measurement is sampled via the following approach:

$$e_{t+1} \sim P'_E(e_{t+1} \mid s_t), \quad s_{t+1} \sim f(e_{t+1}, o_{t+1}). \tag{2}$$

2. **Perception-Aware Quality Estimation:** Evaluate the quality of the candidate measurement using the perception-aware metric, $Q_M(s_{t+1}, o_{t+1})$.
3. **Candidate Sampling (Perception-Agnostic Exploration):** To choose a well-performing sensor configuration o^*_{t+1} for observing a promising state s^*_{t+1}, we first sample k candidate configurations, where k is a hyperparameter controlling the overall processing time. These configurations form a set: $O_{t+1} := \{o^{(1)}_{t+1}, o^{(2)}_{t+1}, \ldots, o^{(k)}_{t+1}\}$, sampled uniformly without replacement from a state-agnostic policy π_{unif} over the sensor configuration space $\mathcal{O}$. For each sampled candidate $o^{(i)}_{t+1}$ ($i = 1, \ldots, k$), we obtain the corresponding candidate state $s^{(i)}_{t+1}$ according to Eq. (2). The set of candidate states is then defined as: $S_{t+1} := \{s^{(1)}_{t+1}, s^{(2)}_{t+1}, \ldots, s^{(k)}_{t+1}\}$.
4. **Optimal Sensing Selection (Perception-Aware Sensing Policy) and State Observation:** Finally, the perception-aware sensing policy π'_{sense} selects the optimal sensing configuration o^*_{t+1} among candidates based on the perception-aware quality metric Q_M: $o^*_{t+1} = \arg\max_{i \in \{1,\ldots,k\}} Q_M(s^{(i)}_{t+1}, o^{(i)}_{t+1}) =: \pi'_{\text{sense}}(O_{t+1} \mid S_{t+1}, Q_M)$.
   The agent then observes the perceptually optimal state using the chosen sensing configuration: $s^*_{t+1} \sim f(e_{t+1}, o^*_{t+1})$.

5.2 Single-Modal Continuous Perception Tasks

As shown in Fig. 3c, we generalize the single-shot adaptive sensing scenario (Lens; Fig. 3b) to a sequential, continuous reinforcement learning (RL) setting without explicit physical actions. We formalize this scenario as a Markov Decision Process (MDP), defined by $\mathcal{M} = (\mathcal{S}, \mathcal{O}, P'_E, Q_M)$, where the agent's decisions exclusively involve adaptively selecting sensing configurations $o_t \in \mathcal{O}$. At each timestep t, the agent performs the following:

1. **State Observation:** Observe current state $s_t \in \mathcal{S}$.
2. **Perception-Based Sensing Selection:** Evaluate perception-aware quality metric $Q_M(s_t, o_t)$ and select the next sensor configuration via policy π_{sense}: $o_{t+1} \sim \pi_{\text{sense}}(o_{t+1} \mid s_t, o_t, Q_M)$.
3. **Environment Response and Sensor Measurement:**

$$e_{t+1} \sim P'_E(e_{t+1} \mid s_t), \quad s_{t+1} \sim f(e_{t+1}, o_{t+1}).$$

In contrast to Lens, where sensing parameters were selected randomly or via simple heuristics, the sensing policy here explicitly maximizes the model-centric quality metric Q. Drawing inspiration from how infants progressively refine gaze control based on continuous perceptual feedback [82, 46], our closed-loop adaptive sensing strategy promotes efficient, coherent sensor exploration, resulting in robust and improved performance in single-modal continuous perception tasks.

5.3 Continuous Sensorimotor Tasks

Key Intuition: Closed-loop Feedback Learning in Humans. In real-world sensorimotor tasks, agents must dynamically co-adapt sensing parameters and physical actions to maintain balance and locomotion across varied terrains. Humans instinctively—or using tools—modulate sensor gains (e.g., foot-pressure thresholds) and adjust actions (e.g., stride lengths) according to environmental conditions such as icy, rocky, or uphill surfaces. Through repeated sensorimotor experiences within this closed-loop feedback framework, perception continuously informs actions, and actions shape subsequent sensory inputs, enabling robust, adaptive behavior.

We formalize this scenario as a Markov Decision Process (MDP) defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P_E, Q_M)$, where the agent jointly selects sensing configurations $o_t \in \mathcal{O}$ and physical actions $a_t \in \mathcal{A}$. At each timestep t, the agent executes:

1. **State Observation:** Observe current state $s_t \in \mathcal{S}$.
2. **Perception-Aware Action Selection:** Select the next action via policy π_{action}, conditioned on perception-aware quality feedback: $a_t \sim \pi_{\text{action}}(a_t \mid s_t, o_t, a_{t-1}, Q_M)$.
3. **Perception-Aware Sensing Selection:** Select the next sensor configuration via policy π_{sense}:
$$o_{t+1} \sim \pi_{\text{sense}}(o_{t+1} \mid s_t, o_t, Q_M).$$
4. **Environment Response and Sensor Measurement:**
$$e_{t+1} \sim P_E(e_{t+1} \mid s_t, a_t), \quad s_{t+1} \sim f(e_{t+1}, o_{t+1}).$$
5. **Reward Collection:** Given a hyperparameter λ, the agent balances task-oriented performance with sensing-quality feedback, enabling efficient joint learning of sensing and action policies: $r_{t+1} \sim R(s_t, a_t, o_t) = R_{\text{task}}(s_t, a_t) + \lambda\, Q_M(s_t, o_t)$. where $R_{task}$is a task-specific reward function that quantifies the effectiveness or success of the agent's actions in achieving the intended physical objective (e.g., duration of maintained balance).

Adaptation to Multi-modality (Sensors) Settings. Beyond adapting sensor parameters within a single modality, agents may also dynamically allocate attention across multiple sensory modalities depending on the task context. For example, when standing still, weight shifts toward the forefoot increase reliance on toe pressure sensors, whereas a lateral push engages ankle proprioceptors more heavily to restore balance. Formally, at each timestep t, we introduce a modality-weight vector $w_t \in \mathbb{R}^N$ across N available sensory modalities (e.g., pressure, torque, IMU), typically normalized so that $\sum_{n=1}^{N} w_t[n] = 1$. Extending the single-modality sensing policy π_{sense} and measurement function f, we define the multi-modality sensing policy $\pi_{\text{multi-sense}}$ and measurement function $f_{\text{multi-sense}}$ as:

$$(o_{t+1}, w_{t+1}) = \pi_{\text{multi-sense}}\big(o_{t+1}, w_{t+1} \mid s_t, o_t, w_t, Q_M\big), \quad s_{t+1} = f_{\text{multi-sense}}\big(s_t, o_t, w_t\big).$$

By adapting w_{t+1} in response to environmental perturbations—such as increasing reliance on ankle sensors when experiencing lateral instability—the agent effectively focuses its sensing resources on the most informative modalities. This mechanism mirrors human sensorimotor reflexes, promoting rapid, robust adaptation and recovery.

Humanoid-Scale Multimodal Tasks under Sparse Rewards. For challenging humanoid tasks—such as opening a bottle cap—agents typically receive a binary, sparse reward $R_{\text{sparse}} \in \{0, 1\}$ only upon successful task completion. To facilitate efficient exploration and learning under these sparse reward conditions, we introduce intermediate perception-aware, continuous sensing-quality metrics Q_i. These metrics quantify intermediate progress or cross-modal sensor feedback, thus providing informative and dense guidance.

For instance, if visual alignment is uncertain while closing a bottle cap, tactile feedback indicating increased grip tightness can enhance the visual sensing-quality metric. Conversely, uncertain tactile sensing can be clarified by visual information regarding precise object positioning. Formally, such cross-modal or intermediate-quality metrics are defined as follows:

$$Q_{\text{grip}}(s_t, a_{t-1}, o_t^{\text{tact}}), \quad Q_{\text{vis}}(s_t, a_{t-1}, o_t^{\text{cam}}, o_t^{\text{tact}}),$$

where Q_{grip} measures grip stability from tactile sensors, and Q_{vis} evaluates visual alignment accuracy informed by both visual and tactile sensor configurations.

These quality metrics can be integrated into a composite reward function:

$$R_t = R_{\text{sparse}} + \lambda_{\text{tact}}\, Q_{\text{grip}}(s_t, a_{t-1}, o_t^{\text{tact}}) + \lambda_{\text{vis}}\, Q_{\text{vis}}(s_t, a_{t-1}, o_t^{\text{cam}}, o_t^{\text{tact}}),$$

where hyperparameters λ_{tact} and λ_{vis} control the relative contribution of each modality. Each metric Q_i explicitly captures task-relevant aspects of sensor data quality, such as grip stability, visual accuracy, uncertainty reduction, or information gain. By leveraging these intermediate, modality-specific metrics, the agent receives rich, continuous feedback tailored to downstream objectives, effectively mitigating the exploration challenges posed by sparse environmental rewards.

6 Challenges and Counterarguments

While adaptive sensing offers transformative potential to reshape AI design and operation, its widespread adoption faces several critical challenges:

- **Lack of Benchmarks:** Existing benchmarks [17, 41, 50, 20] rarely capture sensor and environmental variations. Recent adaptive-sensing benchmarks [7, 6] remain narrowly focused on image classification tasks, limiting generalizability and broader methodological insights.
- **Unspecified Objectives:** Existing adaptive sensing strategies rely on metrics such as model confidence or out-of-distribution (OOD) scores [6], which inadequately capture data-quality under realistic covariate shifts. This hampers accurate assessments and practical applicability.
- **Complexity of Multi-modal Sensor Spaces:** Multi-modal sensors enlarge the perceptual and solution spaces, increasing learning complexity—especially under sparse or delayed reward conditions.
- **Performance Trade-offs:** Adaptive sensing strategies may initially not match peak performance of large-scale, resource-intensive models. Stakeholders may prioritize immediate maximal accuracy over long-term sustainability and efficiency.
- **Integration Barriers with Existing Frameworks:** Implementing adaptive sensing may increase complexity and require significant modifications to existing AI pipelines, software stacks, and hardware platforms [1, 52, 79, 75], posing logistical, economic, and organizational challenges.
- **Ethical and Privacy Concerns:** Real-time, context-aware sensor optimization introduces potential ethical risks related to privacy, particularly when operating in sensitive environments.

7 Open Research Directions

Addressing the identified challenges requires concerted investigation across the following strategic research avenues:

- **Standardized Benchmarks:** Develop comprehensive, realistic benchmarks that evaluate dynamic sensor-level adaptation for various tasks, environments, and modalities, and develop synthetic simulations informed by the real-world data, enabling fair, robust, and comparable assessments.
- **Data-quality Metrics:** Develop test-time evaluation metrics reflecting the model's perception of sensor data quality under covariate shifts, leveraging both aleatoric and epistemic uncertainty. This will enable effective guidance and informed decisions in adaptive sensing.
- **Algorithmic Innovation in Real-time Adaptation:** Advance efficient algorithms using reinforcement learning [68], adaptive control [37], and online learning [27] to optimize sensor parameters in dynamic environments. Developing frameworks to explore and exploit sensor configuration spaces will enhance responsiveness and efficiency.
- **Co-Development of Models and Sensor Strategies:** Develop tightly integrated approaches that simultaneously optimize AI model architectures and adaptive sensor parameters, leveraging mutual feedback for improved generalization and robustness.
- **Multimodal and Language-Driven Adaptive Sensing:** Integrate language-based and multimodal context into sensor adaptation strategies, facilitating intuitive human-AI interaction and broader applicability across domains such as robotics, healthcare, and interactive systems.
- **Privacy-Preserving Sensor Optimization:** Investigate secure, lightweight, and privacy-aware adaptive sensing methods suitable for on-device deployment. Interdisciplinary collaboration involving AI experts, ethicists, and policymakers will be critical for maintaining public trust.

8 Conclusion

In this paper, we have advocated for adaptive sensing as a critical paradigm shift for overcoming fundamental limitations to the prevailing model-centric approach in AI. Drawing inspiration from robust biological sensory systems, adaptive sensing provides a practical pathway toward environmentally sustainable, computationally efficient, ethically responsible, and equitably distributed AI capabilities. By actively optimizing sensor parameters at the input stage, adaptive sensing significantly alleviates the computational and economic burdens currently in large-scale model training and deployment.

We have identified critical gaps in existing methodologies, their limited consideration of continuous, closed-loop interactions essential for real-world embodied agents. Addressing these gaps through targeted research—such as standardized adaptive benchmarks, real-time adaptation algorithms, privacy and ethical standards, and multimodal sensor integration—will be vital. There are urgent, interdisciplinary opportunities that the broader AI community is uniquely positioned to tackle.

Given the escalating environmental, economic, and societal pressures resulting from conventional AI scaling, embracing adaptive sensing is not simply advantageous; it is imperative. We call upon the AI community to prioritize this research agenda. Now is the moment for decisive collective action—by integrating adaptive sensing, we can ensure AI's trajectory aligns fundamentally with ecological responsibility, ethical integrity, and global equity.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Peri Akiva, Matthew Purri, and Matthew Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8203–8215, 2022.

[4] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3):1150–1158, 2018.

[5] Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1453–1463, 2023.

[6] Eunsu Baek, Sung hwan Han, Taesik Gong, and Hyung-Sin Kim. Adaptive camera sensor for vision models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22294–22303, 2024.

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[9] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

[13] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[14] Keylin Alessandra Castro, Joselyn Alvarado Siwady, Erika Castillo, Alberto Alonzo, Manuel Cardona, and María Elena Perdomo. Artificial intelligence for all: Challenges and harnessing opportunities in ai democratization. In *2024 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 143–148. IEEE, 2024.

[15] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[16] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*, 2024.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[19] Kathleen C Fraser and Svetlana Kiritchenko. Examining gender and racial bias in large vision-language models using a novel dataset of parallel images. *arXiv preprint arXiv:2402.05779*, 2024.

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[21] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022.

[22] Yuan Guo, Yun Wang, Qianqian Tong, Boxue Shan, Liwen He, Yuru Zhang, and Dangxiao Wang. Active electronic skin: an interface towards ambient haptic feedback on physical surfaces. *Npj Flexible Electronics*, 8(1):25, 2024.

[23] Vincent Gutschick. Should you use a digital camera in your research? *Bulletin of the ecological society of America*, 83(3):176–180, 2002.

[24] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

[25] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[26] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[27] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[28] Jacqueline Humphries, Kelly O'Brien, Luke O'Brien, and Nathan Dunne. Impact of illuminance on object detection in industrial vision systems using neural networks. In *2020 2nd International Conference on Artificial Intelligence, Robotics and Control*, pages 24–28, 2020.

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, jul 2021. If you use this software, please cite it as below.

[30] Chutian Jiang, Emily Kuang, and Mingming Fan. How can haptic feedback assist people with blind and low vision (blv): A systematic literature review. *ACM Transactions on Accessible Computing*, 18(1):1–57, 2025.

[31] Kyo-Hoon Jin, Kyung-Su Kang, Baek-Kyun Shin, June-Hyoung Kwon, Soo-Jin Jang, Young-Bin Kim, and Han-Guk Ryu. Development of robust detector using the weather deep generative model for outdoor monitoring system. *Expert Systems with Applications*, 234:120984, 2023.

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

[33] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[35] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[36] Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24, 2023.

[37] Ioan Doré Landau, Rogelio Lozano, Mohammed M'Saad, and Alireza Karimi. *Adaptive control: algorithms, analysis and applications*. Springer Science & Business Media, 2011.

[38] Cornelis P Lanting, Paul M Briley, Christian J Sumner, and Katrin Krumbholz. Mechanisms of adaptation in human auditory cortex. *Journal of neurophysiology*, 110(4):973–983, 2013.

[39] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[40] Xingge Li, Shufeng Zhang, Xun Chen, Yashun Wang, Zhengwei Fan, Xiaofei Pang, and Jingwen Hu. Robustness of visual perception system in progressive challenging weather scenarios. *Engineering Applications of Artificial Intelligence*, 119:105740, 2023.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[42] Garrett Manion and Thomas Stokkermans. The effect of pupil size on visual resolution. *StatPearls*, 2024.

[43] Valerio Marsocci, Nicolas Gonthier, Anatol Garioud, Simone Scardapane, and Clément Mallet. Geomultitasknet: Remote sensing unsupervised domain adaptation using geographical coordinates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2075–2085, June 2023.

[44] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[45] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023.

[46] Michiko Miyazaki, Hideyuki Takahashi, Matthias Rolf, Hiroyuki Okada, and Takashi Omori. The image-scratch paradigm: a new paradigm for evaluating infants' motivated gaze control. *Scientific reports*, 4(1):5498, 2014.

[47] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[48] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[50] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[51] Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. Position: A call for embodied AI. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39493–39508. PMLR, 21–27 Jul 2024.

[52] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[53] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

[54] Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, pages 74–83. Springer, 2020.

[55] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[56] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[57] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[59] Gabriele Ruggeri, Debora Nozza, et al. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023.

[60] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[61] Reva Schwartz, Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. *Towards a standard for identifying and managing bias in artificial intelligence*, volume 3. US Department of Commerce, National Institute of Standards and Technology ..., 2022.

[62] Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024.

[63] Muhammad Ali Shah, Ibrar Ali Shah, Duck-Gyu Lee, and Shin Hur. Design approaches of mems microphones for enhanced performance. *Journal of sensors*, 2019(1):9294528, 2019.

[64] Chuanbeibei Shi, Ganghua Lai, Yushu Yu, Mauro Bellone, and Vincezo Lippiello. Real-time multi-modal active vision for object detection on uavs equipped with limited field of view lidar and camera. *IEEE Robotics and Automation Letters*, 8(10):6571–6578, 2023.

[65] Larry Squire, Darwin Berg, Floyd E Bloom, Sascha Du Lac, Anirvan Ghosh, and Nicholas C Spitzer. *Fundamental neuroscience*. Academic press, 2012.

[66] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on robot learning*, pages 477–490. PMLR, 2022.

[67] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020.

[68] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[69] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[70] Brevin Tilmon, Zhanghao Sun, Sanjeev J Koppal, Yicheng Wu, Georgios Evangelidis, Ramzi Zahreddine, Gurunandan Krishnan, Sizhuo Ma, and Jian Wang. Energy-efficient adaptive 3d sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5054–5063, 2023.

[71] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[73] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[74] Loes CJ Van Dam, Myrthe A Plaisier, Catharina Glowania, and Marc O Ernst. Haptic adaptation to slant: No transfer between exploration modes. *Scientific Reports*, 6(1):34412, 2016.

[75] Fernando Van Der Vlist, Anne Helmond, and Fabian Ferrari. Big ai: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, 11(1):20539517241232630, 2024.

[76] Aimee Van Wynsberghe. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218, 2021.

[77] Jun Wang, W-Q Yan, Mohan S Kankanhalli, Ramesh Jain, and Marcel JT Reinders. Adaptive monitoring for video surveillance. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, volume 2, pages 1139–1143. IEEE, 2003.

[78] Michael A Webster, Mohana Kuppuswamy Parthasarathy, Margarita L Zuley, Andriy I Bandos, Lorne Whitehead, and Craig K Abbey. Designing for sensory adaptation: What you see depends on what you've been looking at-recommendations, guidelines and standards should reflect this. *Policy Insights from the Behavioral and Brain Sciences*, 11(1):43–50, 2024.

[79] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[80] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[81] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

[82] Zhengyang Yu, Arthur Aubret, Marcel C Raabe, Jane Yang, Chen Yu, and Jochen Triesch. Active gaze behavior boosts self-supervised object learning. *arXiv preprint arXiv:2411.01969*, 2024.