https://doi.org/10.1007/s11704-026-60382-7

REVIEW ARTICLE

Selective attention or inattention? Enabling efficient perception in embodied intelligence agents

Ruonan XU, Bin GUO✉, Ke MA, Lei WU, Yasan DING, Yao JING, Zhiwen YU

School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Received February 27, 2026; accepted May 12, 2026

E-mail: guob@nwpu.edu.cn

© The Author(s) 2026. This article is published with open access at link.springer.com and journal.hep.com.cn

Abstract

As embodied intelligence increasingly operates in the physical world, perception efficiency, robustness, and adaptability become central bottlenecks separating artificial agents from biological organisms. Unlike conventional perception pipelines, biological perception is inherently selective and resource-aware: attention amplifies task-relevant information, while inattention suppresses distraction and redundancy. This survey investigates attention-driven adaptive perception as a promising mechanism for improving embodied systems. We consolidate interdisciplinary findings from neuroscience, cognitive science, robotics, and machine learning to clarify the distinct yet complementary roles of selective attention and inattention in optimizing perceptual processing. Building on these insights, we introduce the Unified Selective Attention and Inattention Model (U-SAIM) and extend it into a three-tier architectural framework that maps biological principles to embodied perception systems. Guided by this structure, we review existing implementations across robotics, multimodal fusion, lightweight perception, and attention modulation strategies, extracting design patterns and performance implications. Moreover, we discuss the challenges and opportunities for future development. Together, these perspectives chart a unified theoretical and practical roadmap toward attentional embodied perception—agents capable of perceiving, reasoning, and acting with the adaptive intelligence, efficiency, and flexibility that define human perception.

Keywords

adaptive perception; human-inspired cognitive frameworks; attention mechanisms; embodied AI

1 Introduction

Embodied Artificial Intelligence (Embodied AI) represents an emerging paradigm in Artificial Intelligence (AI) that emphasizes the dynamic sensorimotor coupling between an agent's body and the physical environment [1,2]. By grounding intelligence in sensorimotor interactions, it aspires to achieve human-like capabilities through a closed loop of adaptive perception, context-aware reasoning, and goal-directed action [3,4]. Developing such intelligence requires a complex, multidisciplinary effort that involves knowledge and methods from robotics, computer vision, sensor technologies, neuroscience, cognitive science, and related fields. Over the past few years, rapid progress has been made in embodied AI technologies, resulting in agents that can navigate complex environments [5], manipulate objects [6], and interact with humans [7] in increasingly sophisticated ways.

Despite significant advances in embodied AI technologies, a substantial gap remains between embodied AI agents and human agents. In dynamic and unstructured environments, current embodied AI agents are often limited in flexibility, efficiency, and robustness [8–10]. Many approaches rely on static heuristics or task-specific rules, resulting in uniform, unprioritized processing of sensory inputs and inefficient allocation of computational resources. A key distinguishing capability of human agents is **adaptive perception**, *a high-level cognitive function that dynamically filters and prioritizes sensory inputs across spatial, temporal, and semantic dimensions* [11,12]. This ability enables humans to focus on task-relevant information, suppress irrelevant stimuli, and flexibly adapt perception according to changing goals and environmental contexts, achieving both flexibility and efficiency [13,14]. For instance, humans can effortlessly ignore background chatter to focus on a conversation, or rapidly shift visual focus toward a moving object while driving. These capabilities are supported by hierarchical, feedback-rich neural mechanisms that allow perception to be tuned in real-time based on contextual cues, goals and prior experiences [15]. Importantly, human adaptive perception is fundamentally distinct from attention mechanisms commonly used in contemporary computing (e.g., transformers [16]). While computational attention typically operates on preprocessed data at later stages, human attention-based adaptive perception emerges during early sensory acquisition, allowing the nervous system to dynamically filter and

prioritize incoming stimuli. By efficiently allocating perceptual resources and adjusting attention on the fly, humans achieve flexibility and efficiency that current AI systems lack.

To this end, this work aims to incorporate attention-based adaptive perception into embodied AI agents, thereby bridging the gap between artificial and biological systems [17]. Existing surveys on embodied AI have primarily focused on system architectures [18], learning algorithms [19], and task-level applications [20] such as navigation [21], manipulation [22], and human–robot interaction [23]. However, they have largely overlooked the role of perception—the most fundamental and crucial component for building human-like intelligent agents. Inspired by the human attention mechanism, the core of flexible and efficient adaptive perception, this work provides a comprehensive review of attention-based adaptive perception mechanisms by delving into their cognitive and theoretical foundations, surveying current practical implementations in embodied AI agents, and proposing a conceptual framework to further enable AI agents with such adaptive perceptual capabilities.

To systematically introduce the concept of adaptive perception, we initially propose a taxonomy that divides the human attention mechanism into two complementary categories: selective attention, which amplifies relevant signals, and inattention, which deliberately filters out irrelevant input. Importantly, inattention is not perceptual failure, but a resource-efficient strategy shaped by environmental demands and task goals. Both mechanisms are further analyzed by sensory modality (e.g., visual, auditory, tactile) and functional subtypes (e.g., spatial, temporal, semantic), offering a structured foundation for implementing biologically inspired perceptual control.

- **Selective attention** refers to the process by which sensory or computational resources are preferentially allocated to relevant stimuli. It can be subdivided as follows:
  - *Intramodal selective attention*: Operating within a single modality, such as focusing on a moving visual target among static distractors, or attending to one voice in a crowded auditory scene.
  - *Intermodal selective attention*: Coordinating resources across modalities, such as shifting from vision to touch when visual information is degraded or unreliable.
- **Inattention** is the deliberate suppression or omission of processing for non-essential or redundant stimuli. It supports efficiency by reducing overload and conserving computational resources.

To integrate attention-based adaptive perception into embodied AI, we propose a three-tier framework extending the Unified Selective Attention and Inattention Model (U-SAIM). The framework maps human attentional principles to embodied implementations along three dimensions: control drive (stimulus-driven, goal-directed, or history-modulated), operational level (signal acquisition, feature extraction, and decision integration), and loop integration (open-loop, closed-loop, or predictive). By jointly embedding selective attention and inattention, this framework defines a unified design space that captures both information amplification and suppression, providing a foundation for resource-efficient and adaptive embodied perception.

Based on the above conceptual framework, we further examine how existing embodied perception systems have been inspired by, and have begun to implement, elements of adaptive perception. In practice, these efforts manifest at multiple stages of the perceptual pipeline: from signal acquisition mechanisms such as event-driven sensors and foveated cameras, to feature extraction strategies like saliency-guided filtering and predictive coding, and to multimodal integration schemes that exploit intermodal attention for robust fusion. Although often task-specific and limited in scope, these implementations illustrate how selective attention and inattention can be operationalized within embodied architectures to improve efficiency, adaptability, and robustness.

Notably, current research efforts remain far from fully realizing the potential of adaptive perception in embodied agents. Existing implementations guided filtering to multimodal fusion—remain fragmented, highly task-specific, and often lack the contextual flexibility and robustness required for deployment in complex, unstructured environments. To move beyond these limitations, it is necessary to treat attention not as an isolated module but as a unifying principle that governs perception across the signal acquisition, feature extraction, and multimodal integration stages. We therefore outline the key challenges and emerging opportunities for advancing adaptive perception in embodied agents, focusing on five critical dimensions: multimodal structural bionics, multiple attention integration, perception–computing fusion, perception–motion coordination, and perception–morphology coordination.

Overall, this article is organized as follows. Section 2 provides a comprehensive overview of human attention mechanisms that are particularly relevant to adaptive perception. Section 3 presents the U-SAIM model and its three-tier framework, linking human attentional principles to embodied implementations across primary control drive, operational level, and loop integration. In Section 4, we survey recent efforts aimed at integrating attentional mechanisms into embodied AI agents. Section 5 highlights key challenges and identifies emerging opportunities for enabling attention-driven, efficient perception in real-world systems. Finally, Section 6 concludes the paper. Figure 1 illustrates this organization by showing how the review progresses from human attention mechanisms, to the U-SAIM framework, to embodied implementations, and finally to open challenges and opportunities.

2 Attention-based adaptive perception of human agents

Attention mechanisms have played a pivotal role in optimizing sensory processing throughout human evolution [24]. Although the brain constitutes only about 2% of total body mass, it consumes roughly 20% of the body's resting metabolic energy—highlighting the critical need for effective and efficient resource allocation [25]. For instance, during visual search tasks, humans can rapidly focus on a target object in a cluttered scene by suppressing irrelevant visual stimuli—demonstrating how attention conserves cognitive resources while maintaining high perceptual performance. This is largely

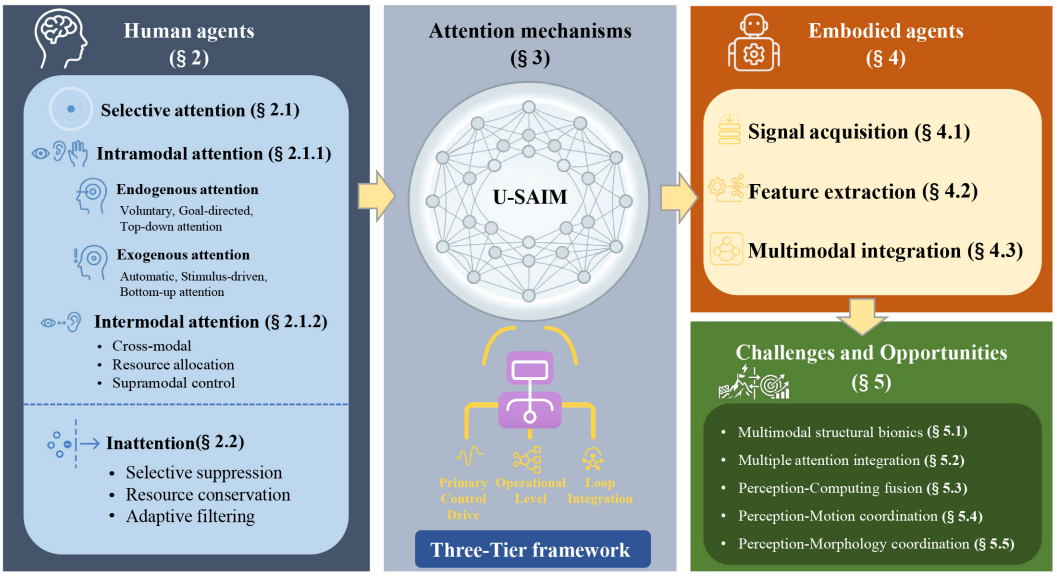


Fig. 1 Overall structure of this review. The discussion progresses from human attention mechanisms in biological agents, to the U-SAIM model and three-tier framework, to embodied implementations organized by signal acquisition, feature extraction, and multimodal integration, and finally to key challenges and opportunities for future research

attributed to adaptive perception, which reduces neural processing demands by selectively enhancing task-relevant inputs while suppressing irrelevant or distracting information.

Initially, we discuss the taxonomy of attention-based adaptive perception mechanisms in biological systems. The key terms and representative phenomena commonly discussed in attention research are summarized in Table 1. Specifically, we categorize these mechanisms based on their functional roles into selective attention and inattention—the former focuses on amplifying task-relevant sensory inputs, while the latter actively filters out irrelevant stimuli. Next, based on modality interaction (e.g., visual, auditory), we further distinguish between intramodal and intermodal attention, depending on whether the process involves single or multiple sensory modalities.

In the following subsections, we provide an in-depth analysis of the objectives, operational principles, and potential inspirations these attentional adaptive mechanisms offer for the design of embodied agents.

2.1 Selective attention

Real-world environments are inherently complex, uncertain, and replete with dynamic, ambiguous stimuli. Within this sensory-rich landscape, selective attention is a core perceptual mechanism that enables the prioritization of behaviorally relevant information while filtering out irrelevant or redundant inputs [36,37].

By dynamically regulating the allocation of limited cognitive and neural resources, selective attention enhances perceptual efficiency, improves decision accuracy, and supports adaptive responses critical for survival [11,14]. It operates across temporal and spatial scales, influencing both early sensory encoding and higher-order integrative processes, and is engaged in tasks ranging from threat detection to complex problem-solving [38,39].

Selective attention is not a unitary construct but manifests in distinct forms depending on the sensory context. Based on the modality in which attentional modulation occurs, it can be classified into intramodal attention—the selective prioritization of information within a single sensory channel—and intermodal attention—the coordination and redistribution of resources across multiple sensory modalities [40,41]. Each type involves distinct but interacting theoretical frameworks and computational models. These mechanisms jointly ensure that perception remains coherent, efficient, and context-sensitive in dynamically changing environments.

2.1.1 Intramodal attention

Similar attention mechanisms operate within single sensory modalities such as vision and audition. In vision, the pop-out effect occurs when objects with prominent features—such as distinctive position, color, shape, orientation, or brightness—immediately capture attention [42,43]. In audition, the cocktail party effect enables the selective following of a specific voice or sound source amid a noisy background [27,44]. Such mechanisms allow attention to be drawn automatically to salient visual or auditory events, including sudden changes [45,46], while also enabling voluntary focus on less conspicuous objects or conversations when task demands require it [12,47]. This dual capacity to shift between stimulus-driven capture and goal-directed selection forms a core characteristic of intramodal attention.

The mainstream view proposes that two complementary pathways exist in the cerebral cortex: the dorsal and ventral systems. The dorsal system, which includes parts of the intraparietal sulcus (IPS) and the frontal eye fields (FEF), is responsible for top-down attention processes guided by goals, prior knowledge, or expectations. The ventral system, which involves the ventral frontal cortex (VFC) and the right temporoparietal junction (TPJ), mediates bottom-up attention processes triggered by sensory inputs or salient and

Table 1 Key terms and phenomena in attention research

Term/Phenomenon	Definition	Category
Pop-out effect [26]	A phenomenon in visual attention where items with highly salient features stand out immediately from their surroundings.	Intramodal attention (Exogenous)
Cocktail party effect [27]	The ability to selectively attend to a single auditory source in a noisy environment.	Intramodal attention
Top-down attention [14]	Goal-directed attentional control guided by internal priorities, expectations, and prior knowledge.	Intramodal attention (Endogenous)
Bottom-up attention [26]	Stimulus-driven attentional control triggered by salient, novel, or unexpected events.	Intramodal attention (Exogenous)
Dorsal frontoparietal network [12]	A brain network supporting top-down attentional control, involving IPS and FEF.	Intramodal attention (Endogenous)
Ventral frontoparietal network [12]	A brain network supporting bottom-up attentional capture, involving VFC and TPJ.	Intramodal attention (Exogenous)
Inhibition of return [28]	A phenomenon where recently attended locations are temporarily suppressed to encourage exploration.	Intramodal attention
Normalization model of attention [29]	A computational model describing attentional modulation as divisive normalization of neural responses.	Intramodal attention
Biased competition hypothesis [14]	A model proposing that multiple stimuli compete for neural representation and attention biases this competition toward task-relevant inputs.	Intermodal attention
Dynamic attentional switching between modalities [30]	The ability to rapidly reallocate attention across sensory modalities in response to changing task demands or environmental cues.	Intermodal attention
Experience-dependent gating [31]	Dynamic regulation of perceptual access based on learned relevance or irrelevance from prior experience.	Intermodal attention / inattention
Selection history effects [32]	Persistent attentional biases toward or away from stimuli based on prior selection experience, independent of current goals or physical salience.	Inattention
Value-modulated attentional capture [33]	Attentional capture driven by stimuli previously associated with high reward.	Inattention
Reward-history-driven attention biases [34]	Sustained attentional priority for stimuli associated with past rewards.	Inattention
Suppression strategies for frequent distractors [35]	Learned inhibition of regularly occurring distractors to improve search efficiency.	Inattention

unexpected stimuli without high-level feedback. When novelty is detected, interactions between the TPJ and IPS can interrupt ongoing top-down control, enabling rapid reorientation to the novel stimulus [12,48]. As illustrated in Fig. 2, this classic top-down/bottom-up control framework explains many cases of selective attention and serves as the foundation for numerous computational models because of its conceptual simplicity and robustness. However, empirical evidence also shows that stimuli lacking both task relevance and salient physical features can still capture attention, indicating the need to extend beyond this traditional dichotomy [32,49]. In the following sections, we introduce endogenous and exogenous attention within this framework, elaborating on additional models, mechanisms, and phenomena that refine and expand the classical perspective.

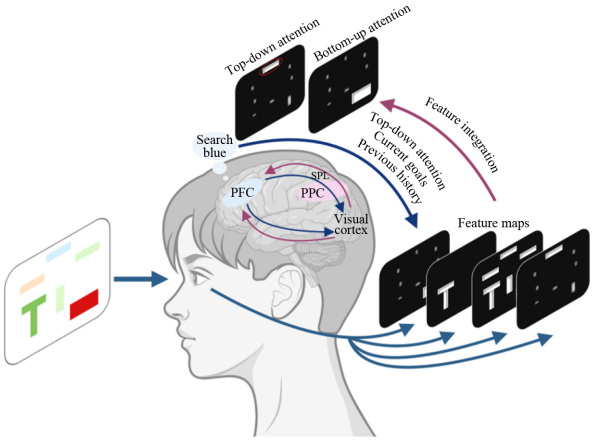


Fig. 2 Classic top-down and bottom-up mechanisms of intramodal visual attention

● **Endogenous attention.** It is also known as voluntary or goal-directed attention, refers to the intentional, top-down modulation of sensory processing within a single modality, guided by internal goals, expectations, or symbolic cues [12,50,51]. This form of attention is anticipatory, temporally sustained, and spatially or feature-selective. It plays a central role in task-driven perception, particularly when sensory inputs are ambiguous, brief, or degraded.

Endogenous attention modulates multiple perceptual dimensions. Spatially, it enhances visual resolution and positional acuity at the attended locus, consistent with findings that attention sharpens psychometric functions and lowers contrast thresholds [52]. In the feature domain, it boosts the gain of channels tuned to specific orientations, directions, or colors, even when these features are not spatially localized [53,54]. These modulations are associated with improvements in discrimination accuracy, faster target detection, and reduced reaction time variability—particularly under conditions of high stimulus uncertainty. Mechanistically, endogenous attention exerts both facilitatory and suppressive influences, forming center-

surround profiles that enhance target representation while inhibiting competing stimuli in adjacent or overlapping feature space [29]. This profile is evident in both psychophysical outcomes and neurophysiological markers, and is interpreted as a form of resource reallocation constrained by neural efficiency.

EEG and ERP studies show that endogenous spatial cues enhance early components (P1, N1) within 100–200 ms of stimulus onset, indicating rapid attentional modulation [55,56]. fMRI and single-unit recordings further reveal that endogenous attention elevates baseline activity and stimulus-evoked responses in modality-specific cortices, sharpening tuning curves and reducing interneuronal variability to improve signal fidelity [57–59]. Beyond sensory encoding, drift-diffusion and sequential sampling models demonstrate that attention accelerates evidence accumulation and lowers decision thresholds [60], while recent work links attentional allocation to both objective accuracy and subjective confidence. Together, these findings highlight endogenous attention's dual role in enhancing perceptual sensitivity and shaping metacognitive evaluation [61].

● **Exogenous attention.** It is also known as automatic or stimulus-driven attention, refers to the rapid, involuntary deployment of perceptual resources toward salient events within a single sensory modality, such as a sudden flash in vision or a sharp onset in audition [51,62]. This bottom-up process is triggered by stimulus salience and operates independently of task goals or volitional control. Its primary function is to prioritize novel or ecologically relevant inputs, thereby ensuring early sensory registration and enabling fast behavioral responses in dynamic environments.

Within a modality, exogenous attention transiently enhances perceptual processing through gain modulation, improving sensitivity and contrast resolution at the cued location or feature. Classic cueing paradigms reveal a biphasic temporal profile: an initial facilitation phase (100–200 ms after stimulus onset) followed by inhibition of return (IOR) [28,63], which reduces the likelihood of reorienting to the same location. This temporal dynamic reflects an adaptive balance between rapid detection and exploratory efficiency. Importantly, attentional capture is not limited to spatial attributes—feature singletons such as abrupt motion or color deviations can also trigger feature-based capture within the same modality, illustrating that intramodal exogenous attention operates flexibly across both spatial and feature dimensions [64,65].

Neurophysiological evidence demonstrates that exogenous attention modulates early sensory stages. EEG and ERP studies report increased amplitudes in early components (e.g., P1 in vision, N1 in audition) [55,66], while fMRI and single-unit data show transient enhancement of stimulus-evoked activity in unimodal cortices [67,68]. These effects occur even without conscious awareness, indicating automatic gain control at early perceptual levels. Furthermore, exogenous attention sharpens sensory tuning functions and reduces trial-to-trial variability (e.g., lower Fano factor and noise correlations), thereby improving signal fidelity [59,69]. Collectively, these findings highlight that exogenous intramodal attention dynamically reallocates resources to salient stimuli, shaping both the speed and quality of sensory processing.

■ 2.1.2 Intermodal attention

It is the process by which cognitive and perceptual systems selectively allocate and coordinate processing resources across different sensory modalities to support coherent perception and adaptive behavior in complex environments, also referred to as intermodal attention [40,70]. This mechanism dynamically regulates the balance between modality-specific sensory gains and integrated, supramodal control, allowing attention to shift toward the modality most relevant to current goals or environmental contingencies.

At the neural level, intermodal attention emerges from the interaction between primary sensory cortices and higher-order supramodal networks in parietal and prefrontal regions [70,71]. These control circuits redistribute resources across modalities, flexibly reweighting sensory evidence depending on stimulus salience, reliability, or task relevance. Electrophysiological evidence highlights the temporal precision of this process: ERP studies show that attentional shifts across modalities modulate early sensory responses within 100 ms of stimulus onset, indicating that intermodal reallocation operates at initial feedforward stages [72]. This rapid responsiveness is particularly crucial in scenarios such as low-visibility environments, where auditory cues become dominant for guiding behavior.

Behavioral and computational models provide converging evidence that intermodal attention shapes decision-making by adjusting the weighting of multisensory evidence. Drift-diffusion and evidence accumulation frameworks reveal that these adjustments optimize both accuracy and speed by modulating decision thresholds and accumulation rates [73,74]. For instance, in audiovisual conflict tasks, attention biases perception toward the more reliable modality, improving task performance [75]. Beyond accuracy, intermodal attention supports perceptual stability, as demonstrated by psychophysical findings where combining auditory and visual cues, such as speech and lip movements, enhances intelligibility in noisy environments [76]. These results underscore its role in maintaining robust perception under uncertainty.

Intermodal attention extends beyond perceptual enhancement to influence higher-level cognition. By reinforcing cross-sensory associations, it facilitates the construction of multimodal representations crucial for learning, memory, and embodied cognition. Supramodal executive networks play a central role in suppressing irrelevant signals while enhancing goal-relevant multisensory cues, thereby reducing cognitive load and stabilizing perceptual experience [77]. Recent EEG findings further demonstrate that attentional dynamics in intermodal contexts track subjective perceptual states rather than purely physical stimulus attributes [78], underscoring their role in conscious perception and adaptive decision-making. Taken together, intermodal attention emerges as a rapid, flexible, and integrative mechanism that ensures efficient allocation of resources across modalities, enabling agents to maintain coherent and adaptive performance in complex environments.

■ 2.2 Inattention

Inattention, often characterized by phenomena such as missed details or inattentional blindness [79,80], is not merely a limitation but a

pivotal adaptive mechanism in human cognition. It functions as a selective suppression process that filters out irrelevant or low-priority stimuli, conserving limited cognitive and neural resources for task-relevant processing. This targeted filtering is indispensable in complex, dynamic environments, where uncontrolled exposure to extraneous input could otherwise overwhelm perceptual systems, impair decision-making, and reduce overall performance.

By attenuating responses to predictable, redundant, or non-salient stimuli, inattention reduces information overload, minimizes cognitive fatigue, and sustains attentional focus. Empirical evidence shows that deliberately ignoring repetitive inputs can enhance working memory, preserve dual-task performance, and improve operational efficiency [81,82]. Beyond efficiency, inattention facilitates divergent thinking by allowing temporary disengagement from immediate tasks, enabling subconscious associative processing. The well-documented "eureka" insights during mind-wandering episodes exemplify how diverting attention from focal tasks can foster novel associations and conceptual breakthroughs [83].

Mechanistically, inattention is supported by pre-attentive filtering that evaluates sensory input below conscious awareness, discarding irrelevant signals before they reach higher-order stages [84]. This process is shaped not only by top-down goals but also by selection history—the persistent biases formed through prior experience. Such history-dependent suppression includes reduced sensitivity to frequent distractors, sensitivity to statistical regularities, and biases associated with previously rewarded or aversive stimuli [32,85]. A central example is value-modulated attentional capture (VMAC): stimuli linked with high reward can continue to attract attention even when counterproductive to current goals [33]. These findings indicate that inattention extends beyond deliberate suppression, encompassing experience-dependent gating mechanisms that reflexively regulate perceptual access.

3 A conceptual framework for attentional embodied perception

Inspired by the attention-based adaptive perception observed in human agents (see Section 2), we propose a conceptual framework to guide the integration of this capability into the design of embodied AI agents. Within this framework, we introduce the concept of *attentional embodied perception*. In contrast to previous embodied perception frameworks, attentional embodied perception is context-sensitive, action-oriented and continuously shaped by the dynamic interplay between sensory inputs, motor actions, and higher-order cognitive processes. Grounded in the attention mechanisms, it mirrors biological perception, where sensory signals are dynamically modulated in response to situational demands, task priorities, and metabolic constraints—thereby optimizing both behavioral performance and resource efficiency of embodied agents.

● From conventional to attentional embodied perception

Perception forms the foundation of embodied perception, integrating exteroceptive modalities (e.g., LiDAR, radar, RGB-D cameras, event-based vision sensors) with interoceptive systems (e.g., IMUs, force-torque sensors, tactile arrays). *Early embodied perception* frameworks rely on *predefined sensor-to-action mappings* and *rigid control pipelines*, which limits adaptability in dynamic environments [1,86]. Despite significant progress in multimodal perception and perception algorithms, embodied agents still face efficiency bottlenecks, often incurring high computational load and energy consumption. Biological systems achieve efficiency by employing predictive coding, adaptive gain control, and context-sensitive filtering to minimize redundant processing [87–89]. However, artificial perception systems frequently lack comparable mechanisms for dynamically balancing accuracy, latency, and resource usage under real-world uncertainty. This gap is especially pronounced in the integration of selective attention into embodied perception, where agents must rapidly reconfigure perceptual focus in response to shifting goals and environmental changes.

Recent developments in *event-driven perception* have shown promise for reducing unnecessary computation by activating processing only when meaningful environmental changes occur [90,91]. Nevertheless, the challenge remains to design robust hybrid attention frameworks that seamlessly transition between local feature extraction and global scene understanding, akin to how humans dynamically adjust attentional scope in response to environmental complexity. Proposed solutions emphasize combining goal-directed, top-down attention with stimulus-driven, bottom-up reactivity, continuously reallocating perceptual resources based on task relevance, uncertainty levels, and contextual cues.

Therefore, existing embodied perception systems still rely on isolated or overly simplified attention implementations, lacking the adaptive modality switching, cross-modal integration, and dynamic suppression strategies characteristic of biological perception. To address this gap, we propose a conceptual framework for attentional embodied perception. To systematically map human attentional embodied perception mechanisms onto embodied systems, we establish the Unified Selective Attention and Inattention Model (U-SAIM), a computationally grounded, biologically inspired formulation that models attention as a multi-drive, resource-constrained optimization problem embedded in the perception–action loop.

3.1 Unified selective attention and inattention model

Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the set of sensory modalities (e.g., vision V, audition A, tactile T). At each time t:

- $x_m(t)$: Raw sensory input from modality m
- $w_m(t) \in [0, 1]$: Attention weight assigned to modality m
- $C_m(w_m)$: Processing cost for modality m, increasing with w_m
- $R_m(w_m, x_m)$: Task-specific information gain (e.g., mutual information, detection accuracy)
- $E(t)$: Total available perceptual energy/computation budget
- $S_m(t)$, $G_m(t)$, $H_m(t)$: **Stimulus-driven**, **Goal-directed**, and **History-modulated** drives for modality m
- $I_m(t)$: Suppression term for inattention, derived from redundancy prediction, noise estimation, or task irrelevance

From a biological perspective, each term in U-SAIM has a direct analogue in the human attention system (HAS). In HAS, $S_m(t)$

reflects **bottom-up salience detection** mediated by sensory cortices and subcortical structures (e.g., superior colliculus), $G_m(t)$ corresponds to **top-down goal modulation** from frontoparietal networks, and $H_m(t)$ embodies **selection history effects** involving learned value and priming. Inattention $I_m(t)$ in HAS is realized through active suppression pathways, such as inhibitory interneuron modulation and thalamic gating, which reduce processing of predictable or irrelevant stimuli.

In embodied attention systems (EAS), these same variables are implemented algorithmically: $S_m(t)$ may be computed via saliency maps from convolutional or transformer architectures; $G_m(t)$ may be derived from task-level planning or reinforcement learning reward functions; $H_m(t)$ can be modeled using statistical priors or adaptive weighting based on interaction history; and $I_m(t)$ can be implemented as channel masking, adaptive sampling rate reduction, or feature pruning. This correspondence ensures that U-SAIM provides a common formalism for analyzing both HAS and EAS under the same resource-constrained control paradigm.

Attention allocation is formulated as

$$\max_{\{w_m(t)\}} \sum_{m\in\mathcal{M}} \left[R_m^{\text{sel}}(w_m, x_m) + R_m^{\text{inatt}}(w_m, x_m)\right] - \lambda \sum_{m\in\mathcal{M}} C_m(w_m)$$
$$\text{s.t.} \sum_m C_m(w_m) \leqslant E(t), \quad w_m(t) \geqslant 0. \tag{1}$$

Here:

- R_m^{sel}: Performance gains from selective enhancement
- R_m^{inatt}: Efficiency gains from suppressing low-value inputs
- λ: Trade-off parameter between performance and cost

This formulation explicitly couples attention as selection with inattention as suppression, aligning with findings that efficient perception arises from both enhancement and filtering.

To connect the static objective in Eq. (1) with online perceptual control, we model the temporal evolution of $w_m(t)$ using a first-order gradient-like dynamics [92,93]:

$$\tau_m \frac{\mathrm{d}w_m}{\mathrm{d}t} = \alpha_m S_m(t) + \beta_m G_m(t) + \delta_m H_m(t) - \eta_m I_m(t) - \gamma_m C'_m(w_m), \tag{2}$$

where

- $\alpha_m S_m(t)$: **Stimulus-driven** salience capture
- $\beta_m G_m(t)$: **Goal-directed** task relevance modulation
- $\delta_m H_m(t)$: **History-modulated** learned prioritization
- $-\eta_m I_m(t)$: Inattention-driven suppression of redundant, noisy, or task-irrelevant input
- $-\gamma_m C'_m(w_m)$: Resource cost feedback under bounded perceptual/computational budget

This law captures **multi-drive attentional control** within a closed sensorimotor loop, enabling both rapid reactive shifts and sustained, strategic allocation. At the level of U-SAIM, the coefficients in Eq. (2) should be understood as balancing terms that regulate the relative influence of stimulus-driven, goal-directed, history-modulated, suppressive, and cost-related factors. In HAS, these influences may arise from interacting neural pathways and physiological constraints, whereas in EAS they may be realized either through manually specified control parameters or through learned weighting, gating, or policy modules. In practice, the update is followed by projection or normalization onto the feasible set defined in Eq. (1), so that the dynamics remains consistent with bounded and resource-constrained attention allocation.

U-SAIM provides a generalizable template: its terms can be implemented via Exogenous saliency maps, reinforcement-learning-based goal relevance estimation, Bayesian history-based weighting, and predictive redundancy suppression. However, applying these equations to real-world embodied agents requires an interpretable structural mapping.

3.2 Three-tier framework for attentional embodied perception

Based on the U-SAIM formulation, attention in embodied perception can be characterized through three orthogonal but interacting dimensions, primary control drive, operational level, and loop integration, each manifesting in both selective attention enhancement and inattention suppression modes. This structure captures how attentional priorities originate, where they act within the perception pipeline, and how they evolve over time.

3.2.1 Primary control drive

The first dimension defines the origin of attentional modulation, determining what triggers the reallocation of perceptual resources. We identify three primary drives, each with distinct mechanisms in HAS and EAS.

● Stimulus-driven control.

In HAS, this form of attention is automatically triggered by salient sensory events such as abrupt motion onsets, sudden luminance changes, or unexpected auditory signals, mediated primarily by ventral attention networks and subcortical structures such as the superior colliculus [94]. It enables rapid orienting without the need for high-level cognitive feedback [95]. In EAS, stimulus-driven control is implemented via saliency detection algorithms—for example, Itti–Koch saliency maps for vision [96], onset detectors for auditory events [97], or LiDAR-based change detection [98]. In selective attention mode, it prioritizes processing of the most salient input channels; in inattention mode, it actively suppresses distracting yet task-irrelevant salience to avoid resource misallocation [35]. The responsiveness of this drive can be further shaped by loop integration mechanisms, where predictive control anticipates salient events before they occur [99].

● Goal-directed control.

In HAS, goal-directed attention is a voluntary, task-driven process that allocates resources according to current objectives, semantic knowledge, or expectations. This mode sustains focus in the presence of competing stimuli and is closely associated with dorsal frontoparietal networks [12]. In EAS, goal-directed control is achieved through task-conditioned perception modules [100], active perception strategies [101], or reinforcement-learning-based

relevance modulation [102]. In selective attention mode, attention is concentrated on task-critical channels [103], while in inattention mode, predictable but irrelevant inputs are suppressed to improve efficiency. Loop integration here often involves closed-loop adjustments, refining attention based on real-time task performance.

● History-modulated control.

In HAS, this drive embodies selection history effects such as value-driven attentional capture, statistical learning, and priming [32,33,104]. These mechanisms can bias attention toward features repeatedly associated with rewards or away from consistently uninformative stimuli [35,105]. In EAS, history-modulated control is realized via adaptive weighting informed by interaction logs [106], Bayesian priors, or learned statistical models of sensory relevance [107]. Selective attention mode reinforces attention toward historically beneficial channels, whereas inattention mode maintains suppression of features or modalities that have shown persistent irrelevance. Predictive integration can leverage history-based priors to preconfigure attention allocation before new data arrive [15].

■ 3.2.2 Operational level

The second dimension specifies where in the perceptual processing hierarchy attentional modulation is applied. In our framework, this dimension is reformulated into three biologically inspired stages of sensory processing—Signal Acquisition, Feature Extraction, and Multimodal Integration—each with direct analogues in both HAS and EAS. Figure 3 compares the corresponding processing layers in human and robotic perception systems and illustrates how selective attention and inattention operate across these layers.

● Signal acquisition layer.

This layer corresponds to the earliest stage of information processing, where raw sensory signals are captured. In HAS, attention at this stage manifests as early sensory gating mechanisms such as retinal adaptation, cochlear gain control, and proprioceptive threshold modulation, which regulate the quantity and quality of incoming data before cortical processing [108,109]. Stimulus-driven control can redirect sensory focus in response to sudden salient events [110], goal-directed control sustains sensor orientation toward task-relevant stimuli [111], and history-modulated control biases sensor deployment according to learned spatial priors or habitual scanning patterns [112]. In EAS, this layer encompasses active camera gaze control, LiDAR beam steering, or adaptive sampling in IMUs, where selective attention mode increases sampling density in high-value regions [113,114], and inattention mode reduces or suspends acquisition from low-value channels to conserve computational and energy resources.

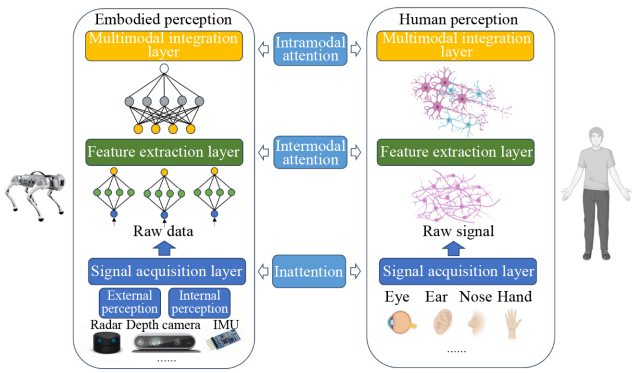


Fig. 3 Comparison of human and robotic perception layers with attention and inattention mechanisms

● Feature extraction layer.

At this stage, raw sensory inputs are transformed into intermediate feature representations. In HAS, attentional modulation appears as spatial attention maps in the visual cortex, feature-selective gain adjustments, or frequency tuning in the auditory system, enabling preferential processing of task-relevant features [115,116]. Stimulus-driven control enhances features with high bottom-up salience [117], goal-directed control amplifies features matching task templates [118], and history-modulated control either reinforces or suppresses features based on prior relevance [119]. In EAS, this layer is implemented as spatial weighting in convolutional feature maps, attention masks in transformer architectures, or feature-channel pruning in deep learning models. Selective attention mode prioritizes discriminative features [120], while inattention mode masks or prunes features associated with noise, redundancy, or consistent irrelevance [121].

● Multimodal Integration layer.

The highest stage involves combining information across multiple sensory modalities. In HAS, attention influences the weighting of different channels during perceptual decision-making [40,122], as in intermodal integration of auditory and visual cues [30], where stimulus-driven control may transiently shift modality dominance, goal-directed control emphasizes the most task-relevant modality [123], and history-modulated control adjusts weights according to prior reliability [124]. In EAS, this corresponds to adaptive multimodal fusion strategies—such as combining camera and LiDAR data in autonomous navigation [125] or audio–vision fusion in human–robot interaction [126]—where selective attention mode boosts the contribution of high-confidence modalities, and inattention mode down-weights or suppresses low-confidence or noisy channels to maintain decision robustness [127].

■ 3.2.3 Loop integration

The third dimension specifies how attention is coupled to the perception–action cycle over time, distinguishing *open-loop*, *closed-loop*, and *predictive* integration. These regimes differ in when and how allocation decisions are made, the degree of feedback available, and the balance between selective enhancement and inattention-driven suppression. A comparison of these three integration modes across human and embodied attention systems is summarized in Table 2.

● Open-loop integration.

Allocation is issued once and not updated within the cycle. In HAS, this corresponds to reflexive, one-shot orienting toward abrupt onsets, often followed by inhibition of return, without subsequent adjustment [28,63,128]. In EAS, it maps to preplanned sensing or

Table 2 Loop integration modes across human attentional systems and embodied attention systems

Mode	Human attentional system	Embodied attention system	Pros / Cons
Open-loop	Reflexive orienting to sudden stimuli; short facilitation then inhibition of return.	Preplanned sensor schedules (fixed foveation policy, static thresholds).	+ Low latency, low compute. – Rigid, poor adaptation.
Closed-loop	Continuous reallocation; smooth pursuit with corollary discharge separating self- vs. external motion.	Active vision with foveated cameras, adaptive beam steering, dynamic exposure/zoom.	+ Adaptive, robust. – Higher compute, possible instability.
Predictive	Anticipatory allocation via forward models and predictive coding.	Model-based scheduling (map-aware LiDAR steering, predictive sensor fusion).	+ Proactive, efficient under uncertainty. – Sensitive to model mismatch.

processing schedules that fix channel priorities or suppressions within a processing window [99]. This mode is computationally frugal but brittle under rapid environmental change.

● **Closed-loop integration**

Attention is continuously reweighted from sensory and motor feedback. In HAS, smooth pursuit eye movements exemplify this process: gaze remains locked onto a moving object by constantly updating allocation based on visual input [129,130]. Similarly, corollary discharge—motor command copies sent to sensory areas—enables the system to distinguish self-induced from external changes, preventing misallocation [131,132]. In EAS, analogous mechanisms appear in real-time sensor control (e.g., active vision or beam steering), where feedback iteratively boosts informative channels and downregulates those whose marginal utility drops, trading higher computational cost for adaptability and stability [101,133].

● **Predictive integration**

Allocation is issued before evidence arrives, informed by internal models. In HAS, forward models and predictive coding support anticipatory activation of expected channels and proactive suppression of predicted noise [87,137,138]. Phenomena such as anticipatory saccades—eye movements directed toward the expected future location of a target—and predictive remapping—shifts in receptive fields before an eye movement occurs—demonstrate how perception is proactively aligned with expected sensory input [139]. In EAS, predictive integration is achieved through model-based controllers, for example preemptively redirecting LiDAR scans toward regions where obstacles are likely to appear while reducing sampling in low-yield areas. This predictive mode ensures timely responses and efficient budget use under uncertainty, complementing the reactive nature of open- and closed-loop control.

Figure 4 illustrates how representative embodied attention systems can be interpreted within the three-dimensional U-SAIM framework. Example (1), active visual search for mobile robots, mainly reflects goal-directed control at the signal acquisition level with strong closed-loop coupling, since sensing is continuously adjusted according to task demands. Example (2), the humanoid robot with crossmodal attention, highlights intermodal coordination at the multimodal integration level, where attentional allocation is shaped by both task relevance and cross-sensory interaction. Examples (3) and (4), which are based on event-driven vision and event-camera perception, are positioned closer to stimulus-driven control and open-loop integration, as they emphasize rapid, low-latency responses to salient environmental changes at early perceptual stages. Together, these examples show how the proposed framework can organize heterogeneous embodied attention systems according to where attentional modulation originates, where it acts, and how it is coupled to the perception–action loop.

4 Enabling technologies for attentional embodied perception

Building on the framework proposed for attentional embodied perception (see Section 3), we aim to map concrete solutions to its key dimensions—namely, primary control drive, operational level, loop configuration, and functional mode. We further examine how these mechanisms address practical challenges in embodied perception, particularly with respect to task performance, efficiency, adaptability, robustness, and real-time deployment. The discussion follows the three operational levels: *Signal Acquisition*, *Feature Extraction*, and *Multimodal Integration*, each of which represents a distinct locus for applying attentional modulation.

In framing this analysis, we recognize a persistent gap between computational implementations and the theoretical constructs of human attention. Certain cognitive theories, while conceptually rich, are highly abstract or even metaphysical, making direct engineering translation infeasible. Nonetheless, these theories provide valuable conceptual anchors for classifying and interpreting existing techniques. Our goal is not to establish absolute one-to-one correspondences, but rather to situate current attention-inspired designs within a broader conceptual and theoretical landscape, clarifying their affinities with established models of selective attention and inattention. Thus, we provide a coherent, system-level perspective on the state of attentional embodied perception, fostering better understanding on how diverse approaches align—or diverge—from biologically grounded principles, and how they are operationalized and advanced across different levels of the perception–action hierarchy.

4.1 Signal acquisition layer

The signal acquisition layer constitutes the first stage of the perception pipeline, where embodied agents gather raw sensory information from the environment. Typical sensor suites include vision (cameras, depth sensors, LiDAR, time-of-flight imagers), audition (microphones, microphone arrays, neuromorphic cochleae), haptics (pressure, force–torque, vibration, temperature sensors), and

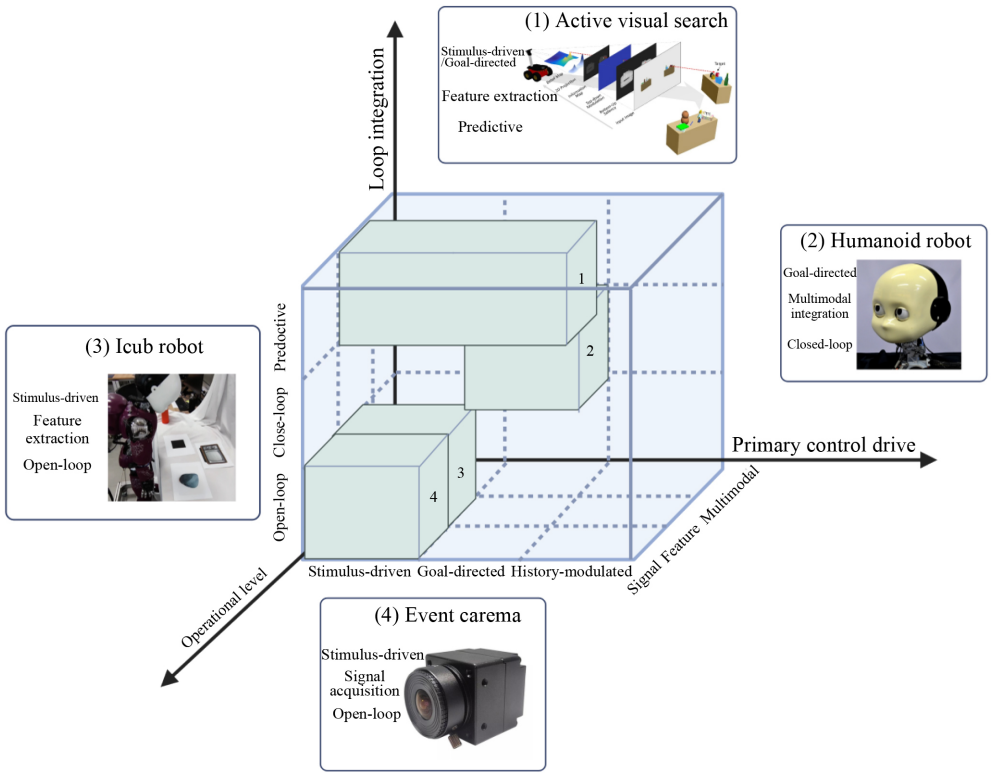


Fig. 4 Representative embodied attention systems mapped within the three-dimensional U-SAIM framework. The examples include: (1) Active visual search for mobile robots [134]. (2) Humanoid robot with crossmodal attention [135]. (3) iCub event-driven vision [136], and (4) event camera-based perception. Each is positioned along the axes of primary control drive, operational level, and loop integration

other modalities such as inertial measurement units (IMUs), proximity sensors, and biosignals (e.g., EMG, EEG). In conventional architectures, each sensor samples its physical signal at fixed temporal intervals, producing a stream of static frames or readings regardless of whether the underlying signal has changed. This uniform sampling introduces two inefficiencies: static or slowly varying signals continue to generate redundant data with no additional information, while highly dynamic signals may be under-sampled, creating a trade-off between sampling rate and bandwidth. Such designs waste computational and energy resources, and can impair responsiveness in fast-changing environments [143,144].

To overcome these limitations, attention-based embodied perception systems adopt an event-driven paradigm in which data is generated only when the signal undergoes a significant change [91]. Representative implementations across visual, auditory, tactile, and olfactory modalities are shown in Fig. 5. Encoding schemes such as pulse-density modulation or sigma–delta modulation enable the event rate to adapt to signal dynamics [145,146]: rapidly changing stimuli generate dense event streams, while static conditions produce little or no output. This principle, inspired by biological encoding in the retina and cochlea, embodies a stimulus-driven selective mechanism that suppresses predictable background activity at the hardware level. In vision, the most mature embodiment of this approach lies in silicon retina technology, which has evolved from mimicking retinal structure to approximating retinal function and even surpassing biological performance [147]. Dynamic Vision Sensors (DVS)

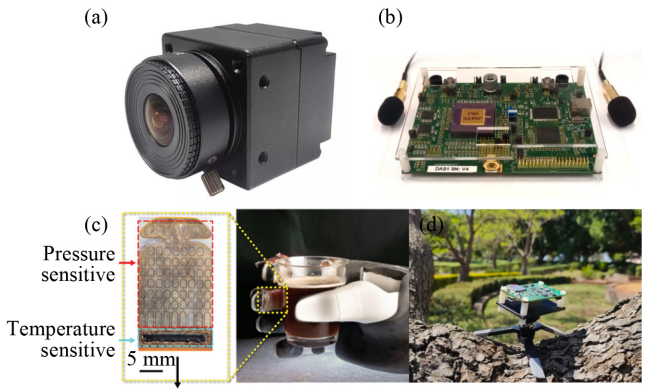


Fig. 5 Event-based neuromorphic sensing across modalities. (a) Event camera (vision); (b) mixed-signal cochlear front-end (audition) [140]; (c) neuromorphic tactile fingertip (touch) [141]; (d) miniaturized electronic nose (olfaction) [142]

[148,149], Asynchronous Temporal Imaging Sensors (ATIS) [90,150], and Dynamic and Active Pixel Vision Sensors (DAVIS) [151] convert logarithmic light-intensity changes into asynchronous spikes, offering ultra-high dynamic range, microsecond latency, and motion-blur-free perception. These sensors, along with integral-sampling designs that mimic the fovea and capture texture features [152], have become central to embodied perception, powering applications from high-speed object tracking to event-based SLAM [146,153].

Building on this foundation, embodied perception is no longer limited to silicon. Artificial synapses implemented with organic and hybrid materials extend event-driven acquisition to diverse modalities [154–156]. Light-sensitive synaptic devices produce postsynaptic currents modulated by wavelength and intensity [157–159], pressure-sensitive synapses provide tactile signals indispensable for manipulation [160–162], sonic synapses emulate binaural auditory processing [163–166], and chemical-sensitive synapses mimic olfaction [167,168] and gustation [169,170]. These devices not only capture stimuli but also encode plasticity, enabling adaptive perception at the hardware level. Complementarily, embodied perception systems combining conventional active sensors with programmable spiking processors such as SpiNNaker [171], Loihi [172], or TrueNorth [173] further enhance acquisition. In such systems, auditory complexes achieve real-time sound localization [174], neuromorphic controllers enable UAV landing through optical flow divergence [175], and hybrid visual systems integrate event cameras with deep learning modules for efficient perception [176].

From the perspective of a unified attention framework, event-driven acquisition realizes selective attention and inattention at the earliest perceptual stage [177,178]. Stimulus-driven spikes ensure responsiveness to novel or transient changes, while adaptive thresholds and predictive modulation enable the system to emphasize task-relevant regions or suppress background activity [179,180]. Moreover, historical regularities can guide predictive sampling, as in adaptive LiDAR scanning that allocates denser beams to regions with frequent dynamic changes while down-sampling static areas [181,182]. Thus, the acquisition layer not only serves as the interface between environment and agent but also as the earliest stage at which attention, inattention, and adaptive resource allocation are instantiated. Its main practical advantage lies in reducing redundant sampling and latency at the source, although its selectivity is often limited at the semantic level unless it is coupled with stronger downstream feature extraction and integration mechanisms.

4.2 Feature extraction layer

The feature extraction layer constitutes the stage in the perception pipeline where raw sensory data are transformed into structured, task-relevant representations. In embodied perception systems, this layer is critical for determining which aspects of the sensory input should be retained, emphasized, or suppressed before passing to higher-level decision modules. From the perspective of our unified attention framework, the feature extraction stage provides multiple opportunities for attentional modulation: stimulus-driven saliency filtering, goal-directed feature reweighting, and history-modulated feature pruning or adaptation. These mechanisms can operate in open-loop configurations, closed-loop feedback systems, or predictive modes, and can function in both selective and inattention roles. Representative technologies at the feature extraction layer, together with their modalities, control drives, loop configurations, and outcomes, are summarized in Table 3.

Bottom-up attention at the feature level prioritizes sensory components that are intrinsically salient according to low-level cues [195] such as contrast, motion [65], color distinctiveness [196],

Table 3 Enabling technologies for attentional embodied perception feature extraction layer

Model	Modality	Sensor	Control drive	Operational level	Loop	Outcome
SNNevProto [183]	Vision	Event camera	Stimulus	Signal acquisition Feature extraction	Open	iCub gaze; ~16 ms to usable saliency; GPU-parity maps.
evProtoDepth [177]	Vision	Event camera	Stimulus	Signal acquisition Feature extraction	Open	Online on iCub; biases gaze to near objects; on-par with frame-based baselines.
SVAM-Net [184]	Vision	RGB camera	Stimulus / Goal	Feature extraction	Open	Competitive on underwater SOD benchmarks; sea trials; real-time feasible.
SP-SLAM [185]	Vision	RGB camera	Stimulus	Feature extraction	Open	Saliency-weighted matching; more robust tracking; steadier mapping.
SalienDet [186]	Vision	RGB camera	Stimulus	Feature extraction	Open	Higher recall for unknown/open-set objects; supports incremental/continual learning.
NAC [187]	Audition	Neuromorphic cochlea	Stimulus	Signal acquisition Feature extraction	Open	Embodied iCub integration; neuromorphic auditory coding.
FISA [188]	Physiology	Electrochemical sweat sensor	Stimulus	Signal acquisition Feature extraction	Open	Real-time perspiration monitoring; multiplexed biochemical and physical sensing.
SE-Net [189]	Vision	/	Goal	Feature extraction	Closed	Channel-wise reweighting; task-relevant feature amplification.
DY-CNN [190]	Vision	/	Goal	Feature extraction	Closed	Input-adaptive kernel selection; efficient task-conditioned representation.
HAT [191]	Vision	/	Stimulus / Goal	Feature extraction	Closed	Top-down & bottom-up scanpath prediction; saliency sequence generation.
RECAP [192]	Vision / Language	/	Goal / History	Feature extraction	Predictive	Resource-efficient finetuning; fewer parameters, minimal loss.
ATP-LLAVA [193]	Vision / Language	/	Goal / History	Feature extraction	Predictive	Token-level predictive inattention; fewer tokens with preserved accuracy.
AdaSpring [194]	Vision / Language	/	Goal / History	Feature extraction	Predictive	Context-adaptive runtime pruning; efficient mobile deployment.

temporal onset [197], and frequency energy distribution [198]. In visual systems, computational saliency models—originating from the Itti–Koch framework [199]—integrate multi-scale feature maps to produce saliency scores that guide downstream processing [195]. In embodied agents, feature saliency maps are often computed directly from event-based sensory streams [90] or from early activations of neural networks, providing a mechanism for pre-attentive suppression of low-salience regions and thereby reducing computational overhead. For vision, event-driven neuromorphic cameras combined with SpiNNaker enable a spiking implementation of the proto-object saliency pipeline for saliency-guided gaze control on the iCub humanoid robot [183]; an event-camera stereo extension (evProtoDepth) produces depth-based proto-object saliency to attract the robot's gaze toward reachable, nearby objects in dynamic 3D scenes [177]. At the same time, in the research of artificial neural networks for embodied perception, many studies on Stimulus-driven attention inspiration have also been conducted [200–202], such as dual-path saliency with a bottom-up branch for underwater robots [184], saliency-guided region selection in visual SLAM [185], and saliency-enhanced detection for autonomous driving [186].

In audition, spectral saliency detection emphasizes transient bursts or unusual harmonic structures, supporting fast localization of novel sounds in cluttered environments [203]. On the iCub Robotic Platform, a neuromorphic architecture encodes spectral and temporal saliency cues to realize stimulus-driven auditory attention, enabling fast scene analysis in parallel to visual saliency maps [187]. In the physiological sensing domain, fully integrated multiplexed sensor arrays for *in situ* perspiration analysis demonstrate how heterogeneous chemical and physical signals can be selectively prioritized for real-time monitoring [188]. Recent advances demonstrate that artificial organic afferent nerves, constructed from organic electrochemical synaptic transistors, can be integrated into embodied agents to realize stimulus-driven prioritization of slip and pressure events, thereby enabling closed-loop feedback for real-time manipulation stability [204].

Comparable principles extend to other modalities: for instance, inertial sensors highlight abrupt motion transitions [205], while olfactory systems may emphasize transient concentration spikes in volatile compounds [206]. Collectively, such feature-level modulation across modalities is inherently stimulus-driven and typically functions in an open-loop selective mode, forming the foundation for fast and lightweight embodied perception.

Top-down attention at the feature level biases representation learning toward task-relevant attributes. In visual pipelines, this can be achieved by dynamically reweighting convolutional feature channels according to task objectives—for example, amplifying texture-sensitive filters in material classification or motion-sensitive filters in navigation tasks. Mechanisms such as squeeze-and-excitation blocks [189], dynamic convolution kernels [190], and attention layers with task embeddings [191] allow embodied agents to condition their feature extraction on high-level goals. In auditory perception, beamforming outputs can feed into adaptive filter banks tuned toward the frequency bands most relevant to the current task (e.g., speech recognition in noisy environments) [207]. In tactile sensing, grip force and contact patch features can be modulated based on the manipulation goal, suppressing non-informative contact data. Goal-directed modulation typically operates in closed-loop mode, allowing real-time feedback to refine feature weighting.

Long-term interaction history provides a basis for predicting which features are likely to be informative in a given context, enabling predictive inattention toward consistently redundant representations. Techniques such as neural network pruning [192], token pruning in Transformer architectures [193], and dynamic subnet selection [194] exploit historical activation statistics to remove or down-weight persistently low-contribution features. In neuromorphic implementations, synaptic depression mechanisms can suppress channels corresponding to predictable inputs [208], while potentiating those associated with novelty or high prediction error. Predictive coding frameworks integrate both goal and history influences [209,210], transmitting only feature-level deviations from an internal model, thereby reducing bandwidth and computation.

In practice, embodied systems often combine these three control drives. For example, a mobile robot performing search-and-rescue may use visual saliency detection (stimulus-driven) to rapidly scan for motion, then apply semantic segmentation masks (goal-directed) to isolate human shapes, while also relying on prior terrain maps (history-modulated) to ignore static background features.

Overall, attentional modulation at the feature extraction level has a more direct impact on downstream perceptual performance, since it determines which representations are amplified, suppressed, or discarded before decision-making. At the same time, the resulting gains in efficiency and adaptability are often accompanied by stronger dependence on model design, task supervision, or learned priors, which may limit generalization across tasks and environments.

4.3 Multimodal integration layer

The multimodal integration layer is responsible for synthesizing heterogeneous sensory streams into a coherent, context-aware representation of the environment. In embodied perception systems, this stage is essential for achieving robustness in scenarios where individual modalities may be incomplete, noisy, or ambiguous. Within our unified attention framework, attentional modulation at the integration stage can be driven by stimulus-driven salience alignment, goal-directed cross-modal weighting, or history-modulated predictive fusion. These controls can operate in open-loop, closed-loop, or predictive configurations, enabling both selective enhancement of informative modalities and suppression of irrelevant or redundant inputs (inattention mode).

In human perception, intermodal exogenous attention allows salient events in one modality to enhance processing in another—for example, a sudden sound directing visual attention toward its source [211]. In embodied systems, this principle is implemented through salience-based gating: high-salience cues from a primary modality trigger increased sampling or processing bandwidth in secondary modalities [99]. For example, in autonomous navigation, LiDAR-detected motion anomalies can cue high-resolution visual capture in the corresponding spatial sector [212], while in human–robot

interaction, abrupt acoustic transients can trigger facial expression or gesture detection pipelines [135]. This mode is typically open-loop and selective, reacting directly to environmental stimuli without extensive task-level reasoning.

Top-down attention in multimodal integration leverages task objectives to dynamically reweight the contribution of each modality. For instance, in low-light environments, autonomous vehicles may increase reliance on LiDAR and radar while down-weighting camera inputs [213,214]; in speech-reading robots, visual lip movement features may be given higher weight when audio signal-to-noise ratio is low [215]. Implementation methods include intermodal attention mechanisms in Transformer-based fusion architectures [216], adaptive mixture-of-experts models that select modality-specific feature extractors [217], and learned gating functions conditioned on task embeddings [218]. This approach operates predominantly in closed-loop mode, where high-level performance feedback informs ongoing reallocation of modality weights.

Long-term interaction history can be exploited to anticipate the reliability and informativeness of different modalities in specific contexts, enabling predictive inattention toward consistently low-value channels. In embodied agents, environmental maps combined with prior sensor reliability statistics can bias fusion toward modalities known to perform well under certain conditions (e.g., radar in fog-prone areas) [219]. In wearable robotics, repeated patterns in physiological sensors can inform predictive suppression of redundant biosignal channels [220]. Predictive coding approaches further reduce intermodal data flow by transmitting only deviations from expected multi-sensory correlations, leveraging prior fusion states to minimize processing demands.

Practical embodied perception systems often blend these control drives. For example, a search-and-rescue robot may use thermal imaging to detect human heat signatures (stimulus-driven), employ mission objectives to prioritize audio–visual confirmation of survivors (goal-directed), and rely on historical terrain–modality performance maps to suppress depth sensing in consistently clear regions (history-modulated). Notable multimodal architectures include vision–LiDAR fusion networks in autonomous driving [221], audio–visual speech recognition systems for assistive embodied agents [222], and vision-tactile integration in grasping [223], where intermodal attention improves grasp stability and object classification.

From a cognitive science perspective, stimulus-driven fusion reflects intermodal exogenous attention, where salient events in one modality reflexively engage others. Goal-directed fusion parallels intermodal endogenous attention, in which higher-order control allocates resources across modalities to serve task goals. History-modulated fusion corresponds to experience-dependent adaptation and statistical learning, where long-term exposure shapes modality prioritization and redundancy suppression. These mechanisms mirror the human brain's integration strategies across primary sensory cortices and supramodal association areas, adapted here into computational architectures for embodied agents. In practical embodied systems, this integration level is particularly important for adaptability and robustness under modality uncertainty or degradation, but its deployment remains constrained by synchronization overhead, fusion complexity, and the difficulty of maintaining stable performance across heterogeneous real-world conditions.

5 Challenges and opportunities

Building on our analysis of selective attention and inattention, and their mapping across signal acquisition, feature extraction, and multimodal integration, we now focus on the key challenges facing efficient attentional embodied perception. Although bio-inspired sensing, computational models, and attentional control architectures have advanced substantially, embodied agents still struggle in unstructured and dynamic environments that demand real-time adaptation, robust decision-making, and resource-efficient operation.

We identify five interrelated research directions: (1) **Multimodal Structural Bionics**, which explores bio-inspired multimodal architectures for robust cross-modal interaction; (2) **Multiple Attention Integration**, which addresses the coordination of spatial, feature-based, and cross-modal attention; (3) **Perception–Computing Fusion**, which seeks joint optimization of perceptual performance and computational efficiency; (4) **Perception–Motion Coordination**, which strengthens bidirectional coupling between perception and motor control; and (5) **Perception–Morphology Coordination**, which examines how physical structure shapes perceptual capability.

Across these directions, another cross-cutting challenge is the lack of consistent empirical protocols for quantifying efficiency–performance trade-offs in heterogeneous attentional embodied systems.

The remainder of this section elaborates on these directions, outlining their technical challenges and potential pathways toward embodied agents with improved adaptability, efficiency, and robustness in real-world settings.

5.1 Multimodal structural bionics

Multimodal structural bionics seeks to create integrated sensory architectures that unify diverse modalities—ranging from biologically inspired event-driven sensors in vision and audition to tactile, proprioceptive, and non-biological sensors such as LiDAR and IMUs—into a coherent framework [224]. While recent advances have produced neuromorphic and event-driven prototypes for individual modalities, most current implementations remain isolated. The key challenge is to transition from fragmented, modality-specific solutions toward large-scale embodied systems where attentional control can operate seamlessly across all sensing channels.

Despite rapid progress, three obstacles limit the realization of fully integrated multimodal systems. First, heterogeneity in sensing principles, data formats, and temporal dynamics hinders direct fusion at the hardware and architectural level [225]. Second, neuromorphic and event-driven designs are typically modality-specific, lacking a unifying standard that would enable consistent attention-driven modulation across modalities [226]. Third, embedding these sensing systems into mobile, morphologically adaptive robots faces constraints in size, power consumption, durability, and real-time

interfacing—making large-scale deployment difficult [227].

Promising research opportunities include the development of unified intermodal sensor platforms in which event-driven encoding is embedded natively across multiple modalities, enabling low-latency and attention-ready signal representation [228,229]. Designing hardware interfaces that directly support attentional functions—such as selective channel activation, dynamic reweighting, and inattention-driven suppression—represents another crucial step [230]. Finally, integrating multimodal sensing with morphological adaptability will allow agents to reconfigure their perceptual apparatus in coordination with structural changes, tailoring perception to environmental and task-specific demands [231,232]. Achieving these goals will require convergence of neuromorphic engineering, robotics hardware design, and attention-modeling research, ultimately providing embodied systems with a scalable substrate for advanced attention-control strategies [233].

■ 5.2 Multiple attention integration

Multiple attention integration refers to the coordinated operation of distinct attentional mechanisms—such as selective attention and inattention—within a unified control architecture [234]. Although embodied perception systems have begun to incorporate these mechanisms [235,236], they are often implemented in isolation, with limited adaptability and minimal interaction across modes. This fragmented approach yields short-term performance gains in specific contexts but also results in redundant computation, inconsistent prioritization, and suboptimal allocation of limited resources when environments change rapidly.

The realization of integrated attention systems is hindered by three primary challenges. First, the absence of a clear control hierarchy leaves unresolved how spatial, feature-based, and intermodal attention should be prioritized and coordinated in real time, particularly under energy and computational constraints [11,237]. Second, high switching latency limits adaptability, as many systems cannot efficiently transition between attentional modes when task demands shift suddenly [238]. Third, the lack of adaptive arbitration mechanisms prevents effective conflict resolution—for instance, when bottom-up saliency highlights one stimulus but top-down task goals emphasize another, leading to inefficiency or instability in perception [12].

Addressing these challenges requires designing unified controllers that govern the interplay of attentional modes. Promising directions include hierarchical attention frameworks that explicitly define roles and priorities across processing stages; predictive scheduling mechanisms that anticipate attentional demands to minimize switching overhead; and resource-aware orchestration strategies that dynamically allocate computational budgets across competing mechanisms. Drawing from biological systems, where attentional processes interact fluidly across modalities and timescales, and from computational control theory, which offers formalisms for arbitration and scheduling, future work aims to establish robust architectures that seamlessly blend diverse attentional modes. Such integration would enable embodied agents to maintain both responsiveness and efficiency in dynamic, resource-constrained environments [239].

■ 5.3 Perception-computing fusion

One of the key directions in advancing efficient embodied perception is perception-computing fusion, which establishes a bidirectional feedback loop between perception and computing. Unlike traditional paradigms where perception and computation are treated as separate processes, this fusion enables adaptive mechanisms that dynamically enhance one capability while the embodied agent performs the other. By seamlessly integrating perception-driven computation and computation-assisted perception, this approach has the potential to significantly improve resource efficiency, generalization, and real-time adaptability, making it a critical pathway for future embodied perception.

To enable perception-computing fusion, two fundamental aspects must be addressed: perception-driven computation and computation-assisted perception. (i) Perception-driven computation focuses on leveraging perception-driven contextual information to optimize computational performance [194]. However, major challenges remain, including contextual representation and mapping, which requires establishing a robust mapping between perceived environmental cues (i.e., device resources [240] and perception data [241]) and computational models to enable adaptive optimization, and online adaptation efficiency [242], which demands computational adjustments with minimal latency and memory cost, particularly in resource-limited embodied systems. (ii) On the other hand, computation-assisted perception integrates computational decision-making to refine perception strategies [243], thereby constructing a unified search space for both perception and computing. A key challenge in this area lies in developing a dynamic framework that optimizes perception strategies—such as sensor morphology [244], active movement [245], and data acquisition parameters [246]—in conjunction with computational adjustments. This line of work focuses on tightly coupling predictive computation with perceptual processes, enabling systems to dynamically adapt their sensing behavior based on computational demands. The ultimate goal is to enhance efficiency, adaptability, and robustness in complex and dynamic environments.

For perception-driven computation, future research should focus on developing adaptive computational models that can efficiently incorporate real-time perceptual feedback. This involves leveraging context-aware learning techniques to enable retraining-free, resource-aware allocation during deployment. For computation-assisted perception, efforts should be directed toward establishing predictable computing demand models that inform dynamic optimization frameworks for perception strategies.

■ 5.4 Perception-motion coordination

Perception-motion coordination is one of the core challenges in the field of embodied intelligence, playing a critical role in enabling agents to interact and navigate efficiently and safely within complex, dynamic environments [247]. This challenge involves not only the synchronization of perception and motion but also their adaptability, ensuring that agents can maintain stable and efficient behavior even when faced with unforeseen environmental changes. Moreover, as artificial intelligence and robotics continue to evolve, this

coordination faces increasingly complex barriers, such as the integration of multi-modal perception systems (e.g., vision, touch, and auditory sensors) with multi-degree-of-freedom (DoF) motion control. One of the key issues in perception-motion coordination is how to seamlessly fuse data from different sensory channels, including vision, touch, and hearing, to create an integrated understanding of the environment [248]. This information then needs to be translated into coherent and precise movement commands. In dynamic environments, where conditions can change rapidly, this process requires real-time updates and adaptations. The perception system must not only have high-precision environmental awareness but also possess the flexibility to adjust perception in response to unexpected or evolving stimuli. This challenge is further intensified by the need for the agent's motion to be fluid and responsive to these changes, ensuring smooth transitions between perception and action without delays or mismatches that could result in task failure.

Recent advancements in adaptive algorithms and deep reinforcement learning have shown promising potential in improving perception-motion coordination [249,250]. For example, DRL frameworks can iteratively optimize an agent's decision-making policy through feedback mechanisms. This approach allows agents to adjust their perception inputs and motion outputs based on trial-and-error, ultimately optimizing their actions for better coordination. Additionally, research into cross-modal learning, which integrates different sensory modalities, is helping to create more robust systems that can maintain stability across a variety of sensory inputs and environmental contexts. Sensor fusion techniques, which combine data from multiple sensors to generate a richer understanding of the environment, are also crucial for realizing effective perception-motion coordination. Despite these advances, several key challenges remain unresolved. A major challenge is achieving efficient, low-latency perception-motion systems that can operate in real-time, particularly in environments that require fast decision-making. How to maintain the synchronization of perception and motion when agents are exposed to continuously changing and unpredictable inputs remains a significant barrier. Furthermore, how these systems can scale to more complex environments, where higher-level cognitive processing may also be required, continues to be an active area of research.

Looking ahead, we argue several promising directions should be explored. These include the development of more robust sensor networks capable of rapidly assimilating data from diverse sensory inputs with minimal delay, the creation of advanced algorithms that can adapt to both environmental changes and task-specific requirements, and the integration of computationally efficient hardware capable of meeting the demands of real-time, high-fidelity decision-making. Additionally, multi-agent coordination could be a growth area, especially for systems that require collective perception and motion coordination in shared environments.

5.5 Perception-morphology coordination

Perception-morphology coordination is a key issue in achieving efficient perception for embodied agents. The morphology of an agent directly affects its perceptual abilities and how it interacts with the environment [251]. The selection, design, and adaptability of an agent's morphology can significantly impact the performance of its perception system, ultimately determining its success in completing tasks and functioning in diverse environments. The challenge in coordinating perception and morphology lies in how to flexibly design and adjust the agent's morphology based on task requirements to maximize perception performance. For example, in different environments, certain configurations of sensors or morphological features may be better suited to specific tasks, such as the field of view of visual sensors or the sensitivity of tactile sensors. In environments where space or sensory needs change dynamically, the agent may need to adjust its morphology to enhance its perceptual capabilities in the most efficient manner. However, as task environments become more complex and unpredictable, agents must possess the ability to rapidly reconfigure both their physical form and perception systems to cope with these evolving demands [252].

The current trend in embodied agent design is to improve perception-morphology coordination through adaptive and flexible morphological adjustments. This has led to the development of reconfigurable robots and morphological computation models [253]. By enabling agents to alter their morphology based on real-time inputs, agents can optimize their sensory input for specific tasks and environmental conditions. For example, robots designed for navigation in confined spaces, such as in disaster recovery scenarios, might change their shape to squeeze through narrow gaps or adjust their sensory configuration to detect subtle environmental signals, like vibrations or changes in air pressure. Perception-morphology coordination also plays a crucial role in reducing the computational burden on the system. By optimizing both the morphology and perception designs, agents can achieve more efficient data collection and processing, which is particularly important when computational power or sensor bandwidth is limited. When these systems work in harmony, they reduce the need for constant reprocessing or redundant sensory data, ultimately improving the agent's overall efficiency.

Another direction for future research lies in exploring optimization strategies for morphology design through evolutionary algorithms and machine learning techniques. These methods can enable agents to achieve more autonomous and intelligent morphological adaptations, where they evolve or learn optimal configurations based on experience, rather than relying on pre-programmed solutions. By training agents with machine learning to adapt to various environmental conditions and task demands, they can develop the ability to dynamically adjust both morphology and perception strategies without human intervention. This would lead to more robust and flexible designs, allowing agents to easily adapt to entirely new environments or cope with environmental changes with minimal effort.

Beyond these five directions, another cross-cutting challenge concerns evaluation. At present, attentional embodied perception systems are usually assessed with task-specific criteria, making cross-system comparison difficult [254]. Future work would benefit from more standardized reporting that jointly considers downstream task performance, computational or energy cost, latency, and the

trade-off between suppression-induced savings and performance degradation [255]. In this sense, the inattention-related gain in U-SAIM should not be understood as a purely abstract quantity, but as an operational measure linked to reductions in redundant sensing or computation under bounded loss in task performance. Related work on perception-latency-aware estimation, event-based low-latency perception, and adaptive perception control suggests that such gains are most meaningfully interpreted through explicit efficiency–performance trade-offs rather than in isolation [176]. Representative empirical indicators may therefore include task success metrics, compute or energy reduction, bandwidth or active-sensor reduction, and normalized trade-off measures that relate resource savings to performance degradation.

6 Conclusion

This work has advanced a unified perspective on attention-based adaptive perception as a foundational principle for achieving human-like intelligence in embodied agents. By uncovering the cognitive mechanisms of the human attention system, we proposed a taxonomy that distinguishes between selective attention—which amplifies relevant stimuli—and inattention—which strategically filters redundant information to enhance efficiency. Building upon these insights, we introduced the Unified Selective Attention and Inattention Model (U-SAIM) and extended it into a three-tier framework that systematically maps human attentional principles onto embodied AI architectures.

Through this framework, perception is no longer treated as a passive sensing process but as an active, resource-aware control system that dynamically regulates information flow across acquisition, computation, and integration. Our review of existing implementations—from event-driven sensors and foveated vision to saliency-guided filtering and multimodal fusion—demonstrates progress toward realizing attentional embodied perception.

However, substantial challenges remain before embodied AI can attain the perceptual adaptability and robustness of biological systems. Future research must integrate attention not as an auxiliary module but as a unifying organizational principle, linking perception with computing, motion, and morphology in closed adaptive loops. By aligning embodied architectures with the principles of selective amplification and strategic suppression, this study provides both a conceptual foundation and a developmental roadmap toward embodied agents capable of perceiving, reasoning, and acting with human-like efficiency and flexibility.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 62532009, 62522215, 62502392, and 625B2150).

Competing interests

Bin GUO is an Editorial Board member of the journal and a co-author of this article. To minimize bias, he was excluded from all editorial decision-making related to the acceptance of this article for publication. The remaining authors declare no conflict of interest.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds

To view a copy of this licence, see website at creativecommons.org/licenses/by/4.0/.

References

[1] Brooks R A. Intelligence without representation. Artificial Intelligence, 1991, 47(1−3): 139−159

[2] Gallagher S. How the Body Shapes the Mind. New York: Clarendon Press, 2006

[3] Cangelosi A, Schlesinger M. Developmental Robotics: from Babies to Robots. Cambridge: MIT Press, 2015

[4] Ravichandar H, Polydoros A S, Chernova S, Billard A. Recent advances in robot learning from demonstration. Annual Review of Control, Robotics, and Autonomous Systems, 2020, 3: 297−330

[5] Awais M, Naseer M, Khan S, Anwer R M, Cholakkal H, Shah M, Yang M H, Khan F S. Foundation models defining a new era in vision: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2245−2264

[6] Ai B, Tian S, Shi H, Wang Y, Pfaff T, Tan C, Christensen H I, Su H, Wu J, Li Y. A review of learning-based dynamics models for robotic manipulation. Science Robotics, 2025, 10(106): eadt1497

[7] Kim C Y, Lee C P, Mutlu B. Understanding large-language model (LLM)-powered human-robot interaction. In: Proceedings of 2024 ACM/IEEE International Conference on Human-Robot Interaction. 2024, 371−380

[8] Tsotsos J K. A Computational Perspective on Visual Attention. Cambridge: MIT Press, 2021

[9] Frintrop S, Rome E, Christensen H I. Computational visual attention systems and their cognitive foundations: a survey. ACM Transactions on Applied Perception (TAP), 2010, 7(1): 6

[10] Putra R V W, Marchisio A, Zayer F, Dias J, Shafique M. Embodied neuromorphic artificial intelligence for robotics: perspectives, challenges, and research development stack. In: Proceedings of the 18th International Conference on Control, Automation, Robotics and Vision. 2024, 612−619

[11] Carrasco M. Visual attention: the past 25 years. Vision Research, 2011, 51(13): 1484−1525

[12] Corbetta M, Shulman G L. Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience, 2002, 3(3): 201−215

[13] Posner M I, Petersen S E. The attention system of the human brain. Annual Review of Neuroscience, 1990, 13: 25−42

[14] Desimone R, Duncan J. Neural mechanisms of selective visual

attention. Annual Review of Neuroscience, 1995, 18: 193−222

[15] Summerfield C, Egner T. Expectation (and attention) in visual cognition. Trends in Cognitive Sciences, 2009, 13(9): 403−409

[16] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 6000−6010

[17] Wang G, Bao H, Liu Q, Zhou T, Wu S, Huang T, Yu Z, Lu C, Gong Y, Zhang Z, He S. Brain-inspired artificial intelligence research: a review. Science China Technological Sciences, 2024, 67(8): 2282−2296

[18] Yan L, Zhao X, Yang B, Wu Y, Dai G, Li J, Tsui C Y, Cheng K T, Zhang Y, Tu F. Robotic computing system and embodied AI evolution: an algorithm-hardware co-design perspective. Journal of Semiconductors, 2025, 46(10): 101201

[19] Liang W, Zhou R, Ma Y, Zhang B, Li S, Liao Y, Kuang P. Large model empowered embodied AI: a survey on decision-making and embodied learning. 2025, arXiv preprint arXiv: 2508.10399

[20] Duan J, Yu S, Tan H L, Zhu H, Tan C. A survey of embodied AI: from simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(2): 230−244

[21] Li B, Han J, Cheng Y, Tan C, Qi P, Zhang J, Li X. Object goal navigation in eobodied AI: a survey. In: Proceedings of the 4th International Conference on Video, Signal and Image Processing. 2022, 87−92

[22] Li G, Wang R, Xu P, Ye Q, Chen J. The developments and challenges toward dexterous and embodied robotic manipulation: a survey. IEEE Robotics & Automation Magazine, 2026, 33(1): 24−38

[23] Sheridan T B. Human-robot interaction: status and challenges. Human Factors: The Journal of the Human Factors and Ergonomics Society, 2016, 58(4): 525−532

[24] Todd J T, Van Gelder P. Implications of a transient-sustained dichotomy for the measurement of human performance. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5(4): 625−638

[25] Wong-Riley M. Energy metabolism of the visual system. Eye and Brain, 2010, 2: 99−116

[26] Treisman A M, Gelade G. A feature-integration theory of attention. Cognitive Psychology, 1980, 12(1): 97−136

[27] Cherry E C. Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, 1953, 25(5): 975−979

[28] Posner M I, Cohen Y. Components of visual orienting. Attention and Performance X: Control of Language Processes, 1984, 32: 531−556

[29] Reynolds J H, Heeger D J. The normalization model of attention. Neuron, 2009, 61(2): 168−185

[30] Spence C, Driver J. Audiovisual links in endogenous covert spatial attention. Journal of Experimental Psychology: Human Perception and Performance, 1996, 22(4): 1005−1030

[31] Bao S, Chang E F, Woods J, Merzenich M M. Temporal plasticity in the primary auditory cortex induced by operant perceptual learning. Nature Neuroscience, 2004, 7(9): 974−981

[32] Awh E, Belopolsky A V, Theeuwes J. Top-down versus bottom-up attentional control: a failed theoretical dichotomy. Trends in Cognitive Sciences, 2012, 16(8): 437−443

[33] Anderson B A, Laurent P A, Yantis S. Value-driven attentional capture. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(25): 10367−10371

[34] Anderson B A. A value-driven mechanism of attentional selection. Journal of Vision, 2013, 13(3): 7

[35] Wang B, Theeuwes J. Statistical regularities modulate attentional capture. Journal of Experimental Psychology: Human Perception and Performance, 2018, 44(1): 13−17

[36] Tsotsos J K. Analyzing vision at the complexity level. Behavioral and Brain Sciences, 1990, 13(3): 423−445

[37] Broadbent D E. Perception and Communication. Amsterdam: Elsevier, 2013

[38] Fiebelkorn I C, Kastner S. A rhythmic theory of attention. Trends in Cognitive Sciences, 2019, 23(2): 87−101

[39] van Moorselaar D, Slagter H A. Inhibition in selective attention. Annals of the New York Academy of Sciences, 2020, 1464(1): 204−221

[40] Talsma D, Senkowski D, Soto-Faraco S, Woldorff M G. The multifaceted interplay between attention and multisensory integration. Trends in Cognitive Sciences, 2010, 14(9): 400−410

[41] Van der Burg E, Olivers C N L, Bronkhorst A W, Theeuwes J. Pip and pop: nonspatial auditory signals improve spatial visual search. Journal of Experimental Psychology: Human Perception and Performance, 2008, 34(5): 1053−1065

[42] Itti L, Baldi P. Bayesian surprise attracts human attention. Vision Research, 2009, 49(10): 1295−1306

[43] VanRullen R. Visual saliency and spike timing in the ventral visual pathway. Journal of Physiology-Paris, 2003, 97(2-3): 365−377

[44] Bronkhorst A W. The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acta Acustica United with Acustica, 2000, 86(1): 117−128

[45] Yantis S, Jonides J. Abrupt visual onsets and selective attention: evidence from visual search. Journal of Experimental Psychology: Human Perception and Performance, 1984, 10(5): 601−621

[46] Näätänen R. Attention and Brain Function. London: Routledge, 1992

[47] Fritz J B, Elhilali M, David S V, Shamma S A. Auditory attention—focusing the searchlight on sound. Current Opinion in Neurobiology, 2007, 17(4): 437−455

[48] Vossel S, Geng J J, Fink G R. Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. The Neuroscientist, 2014, 20(2): 150−159

[49] Egeth H E, Yantis S. VISUAL ATTENTION: control, representation, and time course. Annual Review of Psychology, 1997, 48: 269−297

[50] Posner M I, Snyder C R R. Attention and cognitive control 1. In: Solso R L, ed. Information Processing and Cognition. London: Routledge, 1975, 55−85

[51] Styles E A. The Psychology of Attention. 2nd ed. London: Psychology Press, 2006

[52] Carrasco M, Ling S, Read S. Attention alters appearance. Nature Neuroscience, 2004, 7(3): 308−313

[53] Treue S, Trujillo J C M. Feature-based attention influences motion processing gain in macaque visual cortex. Nature, 1999, 399(6736): 575−579

[54] Maunsell J H R, Treue S. Feature-based attention in visual cortex.

Trends in Neurosciences, 2006, 29(6): 317−322
[55] Mangun G R, Hillyard S A. Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. Journal of Experimental Psychology: Human Perception and Performance, 1991, 17(4): 1057−1074
[56] Hillyard S A, Vogel E K, Luck S J. Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philosophical Transactions of the Royal Society B: Biological Sciences, 1998, 353(1373): 1257−1270
[57] Luck S J, Chelazzi L, Hillyard S A, Desimone R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. Journal of Neurophysiology, 1997, 77(1): 24−42
[58] Mitchell J F, Sundberg K A, Reynolds J H. Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron, 2009, 63(6): 879−888
[59] Cohen M R, Maunsell J H R. Attention improves performance primarily by reducing interneuronal correlations. Nature Neuroscience, 2009, 12(12): 1594−1600
[60] Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural Computation, 2008, 20(4): 873−922
[61] Zizlsperger L, Sauvigny T, Haarmeier T. Selective attention increases choice certainty in human decision making. PLoS One, 2012, 7(7): eadt1497
[62] Smith D T, Rorden C, Jackson S R. Exogenous orienting of attention depends upon the ability to execute eye movements. Current Biology, 2004, 14(9): 792−795
[63] Klein R M. Inhibition of return. Trends in Cognitive Sciences, 2000, 4(4): 138−147
[64] Theeuwes J. Perceptual selectivity for color and form. Perception & Psychophysics, 1992, 51(6): 599−606
[65] Nothdurft H C. Attention shifts to salient targets. Vision Research, 2002, 42(10): 1287−1306
[66] Näätänen R, Picton T. The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. Psychophysiology, 1987, 24(4): 375−425
[67] Ungerleider S K L G. Mechanisms of visual attention in the human cortex. Annual Review of Neuroscience, 2000, 23: 315−341
[68] Busse L, Katzner S, Treue S. Temporal dynamics of neuronal modulation during exogenous and endogenous shifts of visual attention in macaque area MT. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(42): 16380−16385
[69] Reynolds J H, Pasternak T, Desimone R. Attention increases sensitivity of V4 neurons. Neuron, 2000, 26(3): 703−714
[70] Macaluso E, Noppeney U, Talsma D, Vercillo T, Hartcher-O'Brien J, Adam R. The curious incident of attention in multisensory integration: bottom-up *vs*. top-down. Multisensory Research, 2016, 29(6-7): 557−583
[71] Cappe C, Barone P. Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. European Journal of Neuroscience, 2005, 22(11): 2886−2902
[72] Foxe J J, Schroeder C E. The case for feedforward multisensory convergence during early cortical processing. Neuroreport, 2005, 16(5): 419−423
[73] Colonius H, Diederich A. Multisensory interaction in saccadic reaction time: a time-window-of-integration model. Journal of Cognitive Neuroscience, 2004, 16(6): 1000−1009
[74] Kayser C, Shams L. Multisensory causal inference in the brain. PLoS Biology, 2015, 13(2): e1002075
[75] van Atteveldt N M, Formisano E, Goebel R, Blomert L. Top-down task effects overrule automatic multisensory responses to letter-sound pairs in auditory association cortex. Neuroimage, 2007, 36(4): 1345−1360
[76] Alsius A, Navarra J, Campbell R, Soto-Faraco S. Audiovisual integration of speech falters under high attention demands. Current Biology, 2005, 15(9): 839−843
[77] Noppeney U, Ostwald D, Werner S. Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. The Journal of Neuroscience, 2010, 30(21): 7434−7446
[78] Senkowski D, Schneider T R, Foxe J J, Engel A K. Crossmodal binding through neural coherence: implications for multisensory processing. Trends in Neurosciences, 2008, 31(8): 401−409
[79] Simons D J. To see but not to see. Review of Inattentional Blindness by A. Mack and I. Rock (1998). Journal of Mathematical Psychology, 1999, 43(1): 165−171
[80] Simons D J, Chabris C F. Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception, 1999, 28(9): 1059−1074
[81] Lavie N. Distracted and confused?: selective attention under load. Trends in Cognitive Sciences, 2005, 9(2): 75−82
[82] Forster S, Lavie N. Harnessing the wandering mind: the role of perceptual load. Cognition, 2009, 111(3): 345−355
[83] Smallwood J, Schooler J W. The science of mind wandering: empirically navigating the stream of consciousness. Annual Review of Psychology, 2015, 66: 487−518
[84] Luck S J, Gold J M. The construct of attention in schizophrenia. Biological Psychiatry, 2008, 64(1): 34−39
[85] Chelazzi L, Marini F, Pascucci D, Turatto M. Getting rid of visual distractors: the why, when, how, and where. Current Opinion in Psychology, 2019, 29: 135−147
[86] Mataric M J. Behaviour-based control: examples from navigation, learning, and group behaviour. Journal of Experimental & Theoretical Artificial Intelligence, 1997, 9(2-3): 323−336
[87] Friston K. A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences, 2005, 360(1456): 815−836
[88] Carandini M, Heeger D J. Normalization as a canonical neural computation. Nature Reviews Neuroscience, 2012, 13(1): 51−62
[89] Summerfield C, de Lange F P. Expectation in perceptual decision making: neural and computational mechanisms. Nature Reviews Neuroscience, 2014, 15(11): 745−756
[90] Posch C, Matolin D, Wohlgenannt R. A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS. IEEE Journal of Solid-State Circuits, 2011, 46(1): 259−275
[91] Gallego G, Delbrück T, Orchard G, Bartolozzi C, Taba B, Censi A, Leutenegger S, Davison A J, Conradt J, Daniilidis K, Scaramuzza D. Event-based vision: a survey. IEEE Transactions on Pattern Analysis and

Machine Intelligence, 2022, 44(1): 154−180

[92] Allibhoy A, Cortés J. Control-barrier-function-based design of gradient flows for constrained nonlinear programming. IEEE Transactions on Automatic Control, 2024, 69(6): 3499−3514

[93] Schöner G. Dynamical systems approaches to cognition. In: Sun R, ed. The Cambridge Handbook of Computational Psychology. Cambridge: Cambridge University Press, 2008

[94] Alves P N, Forkel S J, Corbetta M, de Schotten M T. The subcortical and neurochemical organization of the ventral and dorsal attention networks. Communications Biology, 2022, 5(1): 1343

[95] Herman J P, Arcizet F, Krauzlis R J. Attention-related modulation of caudate neurons depends on superior colliculus activity. eLife, 2020, 9: e53998

[96] Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 2000, 40(10-12): 1489−1506

[97] Bello J P, Daudet L, Abdallah S, Duxbury C, Davies M, Sandler M B. A tutorial on onset detection in music signals. IEEE Transactions on Speech and Audio Processing, 2005, 13(5): 1035−1047

[98] Stilla U, Xu Y. Change detection of urban objects using 3D point clouds: a review. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 197: 228−255

[99] Bajcsy R, Aloimonos Y, Tsotsos J K. Revisiting active perception. Autonomous Robots, 2018, 42(2): 177−196

[100] Marza P, Matignon L, Simonin O, Wolf C. Task-conditioned adaptation of visual features in multi-task policy learning. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 17847−17856

[101] Aloimonos J, Weiss I, Bandyopadhyay A. Active vision. International Journal of Computer Vision, 1988, 1(4): 333−356

[102] Al-Hamadani M N A, Fadhel M A, Alzubaidi L, Harangi B. Reinforcement learning algorithms and applications in healthcare and robotics: a comprehensive and systematic review. Sensors, 2024, 24(8): 2461

[103] Seminara L, Gastaldo P, Watt S J, Valyear K F, Zuher F, Mastrogiovanni F. Active haptic perception in robots: a review. Frontiers in Neurorobotics, 2019, 13: 53

[104] Chelazzi L, Perlato A, Santandrea E, Della Libera C. Rewards teach visual selective attention. Vision Research, 2013, 85: 58−72

[105] Failing M, Theeuwes J. Selection history: how reward modulates selectivity of visual attention. Psychonomic Bulletin & Review, 2018, 25(2): 514−538

[106] Krause A, Guestrin C. Near-optimal observation selection using submodular functions. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence. 2007, 1650−1654

[107] Taylor A T, Berrueta T A, Murphey T D. Active learning in robotics: a review of control principles. Mechatronics, 2021, 77: 102576

[108] Jarsky T, Cembrowski M, Logan S M, Kath W L, Riecke H, Demb J B, Singer J H. A synaptic mechanism for retinal adaptation to luminance and contrast. Journal of Neuroscience, 2011, 31(30): 11003−11015

[109] Masland R H. Sensory systems: fine-tuning the visual scene. Current Biology, 2005, 15(19): R808−R810

[110] Solomon S G, Kohn A. Moving sensory adaptation beyond suppressive effects in single neurons. Current Biology, 2014, 24(20): R1012−R1022

[111] Laughlin S B. The role of sensory adaptation in the retina. Journal of Experimental Biology, 1989, 146(1): 39−62

[112] Berlyne D E. Novelty and attention: controls for retinal adaptation and for stimulus-response specificity. Psychonomic Science, 1971, 25(6): 349−351

[113] Tilmon B, Koppal S J. SaccadeCam: adaptive visual attention for monocular depth sensing. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. 2021, 6009−6018

[114] Odira J K, Andreff N, Petiet L, Byiringiro J B. Visual servoing of a laser beam through a mirror. Robotica, 2022, 40(9): 3157−3177

[115] Itthipuripat S, Garcia J O, Rungratsameetaweemana N, Sprague T C, Serences J T. Changing the spatial scope of attention alters patterns of neural gain in human cortex. The Journal of Neuroscience, 2014, 34(1): 112−123

[116] Ahveninen J, Hämäläinen M, Jääskeläinen I P, Ahlfors S P, Huang S, Lin F H, Raij T, Sams M, Vasios C E, Belliveau J W. Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(10): 4182−4187

[117] Goddard E, Carlson T A, Woolgar A. Spatial and feature-selective attention have distinct, interacting effects on population-level tuning. Journal of Cognitive Neuroscience, 2022, 34(2): 290−312

[118] Tünçok E, Carrasco M, Winawer J. Spatial attention selectively alters visual cortical representation during target anticipation. Nature Communications, 2025, 16(1): 8746

[119] van Es D M, Theeuwes J, Knapen T. Spatial sampling in human visual cortex is modulated by both spatial and feature-based attention. eLife, 2018, 7: e36928

[120] Cheng H, Zhang M, Shi J Q. A survey on deep neural network pruning: taxonomy, comparison, analysis, and recommendations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10558−10578

[121] Yuan T, Li Z, Liu B, Tang Y, Liu Y. ARPruning: an automatic channel pruning based on attention map ranking. Neural Networks, 2024, 174: 106220

[122] Macaluso E, Driver J. Multisensory spatial interactions: a window onto functional integration in the human brain. Trends in Neurosciences, 2005, 28(5): 264−271

[123] Mozolic J L, Joyner D, Hugenschmidt C E, Peiffer A M, Kraft R A, Maldjian J A, Laurienti P J. Cross-modal deactivations during modality-specific selective attention. BMC Neurology, 2008, 8(1): 35

[124] Körding K P, Beierholm U, Ma W J, Quartz S, Tenenbaum J B, Shams L. Causal inference in multisensory perception. PLoS One, 2007, 2(9): e943

[125] Bai X, Hu Z, Zhu X, Huang Q, Chen Y, Fu H, Tai C L. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, 1090−1099

[126] Xu D, Li H, Wang Q, Song Z, Chen L, Deng H. M2DA: multi-modal fusion transformer incorporating driver attention for autonomous driving. 2024, arXiv preprint arXiv: 2403.12552

[127] Zhang Y, Tu C, Gao K, Wang L. Multisensor information fusion:

future of environmental perception in intelligent vehicles. Journal of Intelligent and Connected Vehicles, 2024, 7(3): 163−176

[128] Taylor T L, Klein R M. On the causes and effects of inhibition of return. Psychonomic Bulletin & Review, 1998, 5(4): 625−643

[129] Krauzlis R J. Recasting the smooth pursuit eye movement system. Journal of Neurophysiology, 2004, 91(2): 591−603

[130] Lisberger S G. Visual guidance of smooth pursuit eye movements. Annual Review of Vision Science, 2015, 1: 447−468

[131] Sommer M A, Wurtz R H. Visual perception and corollary discharge. Perception, 2008, 37(3): 408−418

[132] Crapse T B, Sommer M A. Corollary discharge circuits in the primate brain. Current Opinion in Neurobiology, 2008, 18(6): 552−557

[133] Chen Y, Wang D, Thomas L, Dantu K, Koppal S J. Design of an adaptive lightweight LiDAR to decouple robot-camera geometry. IEEE Transactions on Robotics, 2024, 40: 2254−2271

[134] Rasouli A, Lanillos P, Cheng G, Tsotsos J K. Attention-based active visual search for mobile robots. Autonomous Robots, 2020, 44(2): 131−146

[135] Fu D, Abawi F, Carneiro H, Kerzel M, Chen Z, Strahl E, Liu X, Wermter S. A trained humanoid robot can perform human-like crossmodal social attention and conflict resolution. International Journal of Social Robotics, 2023, 15(8): 1325−1340

[136] D'Angelo G, Voto S, Iacono M, Glover A, Niebur E, Bartolozzi C. Event-driven figure-ground organisation model for the humanoid robot iCub. Nature Communications, 2025, 16(1): 1874

[137] Rao R P N, Ballard D H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 1999, 2(1): 79−87

[138] Keller G B, Mrsic-Flogel T D. Predictive processing: a canonical cortical computation. Neuron, 2018, 100(2): 424−435

[139] Duhamel J R, Colby C L, Goldberg M E. The updating of the representation of visual space in parietal cortex by intended eye movements. Science, 1992, 255(5040): 90−92

[140] Liu S C, van Schaik A, Minch B A, Delbruck T. Asynchronous binaural spatial audition sensor with 2×64×4 channel output. IEEE Transactions on Biomedical Circuits and Systems, 2014, 8(4): 453−464

[141] Lee W W, Tan Y J, Yao H, Li S, See H H, Hon M, Ng K A, Xiong B, Ho J S, Tee B C K. A neuro-inspired artificial peripheral nervous system for scalable electronic skins. Science Robotics, 2019, 4(32): eaax2198

[142] Dennler N, Drix D, Warner T P A, Rastogi S, Casa C D, Ackels T, Schaefer A T, van Schaik A, Schmuker M. High-speed odor sensing using miniaturized electronic nose. Science Advances, 2024, 10(45): eadp1764

[143] Li J, Tian Y. Recent advances in neuromorphic vision sensors: a survey. Chinese Journal of Computers, 2021, 44(6): 1258−1286

[144] Kim Y, Lee C W, Jang H W. Neuromorphic hardware for artificial sensory systems: a review. Journal of Electronic Materials, 2025, 54(5): 3609−3650

[145] Lazar A A, Tóth L T. Perfect recovery and sensitivity analysis of time encoded bandlimited signals. IEEE Transactions on Circuits and Systems I: Regular Papers, 2004, 51(10): 2060−2073

[146] Singh Alvarado A, Rastogi M, Harris J G, Príncipe J C. The integrate-and-fire sampler: a special type of asynchronous Σ − Δ modulator. In: Proceedings of 2011 IEEE International Symposium of Circuits and Systems. 2011, 2031−2034

[147] Liu S C, Delbruck T. Neuromorphic sensory systems. Current Opinion in Neurobiology, 2010, 20(3): 288−295

[148] Lichtsteiner P, Posch C, Delbruck T. A 128×128 120 dB 15μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits, 2008, 43(2): 566−576

[149] Posch C, Serrano-Gotarredona T, Linares-Barranco B, Delbruck T. Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. Proceedings of the IEEE, 2014, 102(10): 1470−1484

[150] Posch C, Matolin D, Wohlgenannt R. An asynchronous time-based image sensor. In: Proceedings of 2008 IEEE International Symposium on Circuits and Systems. 2008, 2130−2133

[151] Brandli C, Berner R, Yang M, Liu S C, Delbruck T. A 240×180 130 dB 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits, 2014, 49(10): 2333−2341

[152] Culurciello E, Etienne-Cummings R, Boahen K A. A biomorphic digital image sensor. IEEE Journal of Solid-state Circuits, 2003, 38(2): 281−294

[153] Guo Y, Hu L, Bai Y, Yao J, Ma L, Huang T. SpikeGS: reconstruct 3D scene captured by a fast-moving bio-inspired camera. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. 2025, 3293−3301

[154] Chan V, Liu S C, van Schaik A. AER EAR: a matched silicon cochlea pair with address event representation interface. IEEE Transactions on Circuits and Systems I: Regular Papers, 2007, 54(1): 48−59

[155] Ved K, Lenk C, Ivanov T, Hövel P, Ziegler M. Bio-inspired, adaptive acoustic sensor: sensing properties in dependence of feedback parameters. AIP Conference Proceedings, 2024, 3062(1): 040011

[156] Yu H, Zhu Y, Zhu L, Lin X, Wan Q. Recent advances in transistor-based bionic perceptual devices for artificial sensory systems. Frontiers in Nanotechnology, 2022, 4: 954165

[157] Wang Y, Yin L, Huang W, Li Y, Huang S, Zhu Y, Yang D, Pi X. Optoelectronic synaptic devices for neuromorphic computing. Advanced Intelligent Systems, 2021, 3(1): 2000099

[158] Lei P, He Z, Cao L, Zhang M, Zheng C, Lin H, Du X, Tao S. Multifunctional organic artificial optoelectronic synapses for neuromorphic computing and a weak-light-sensitive visual system. Journal of Materials Chemistry C, 2025, 13(23): 11707−11717

[159] Choi J H, An H M, Park H L. Polarized light-sensitive optoelectronic synapses for expanding artificial vision systems. Flexible and Printed Electronics, 2025, 10(1): 013501

[160] Lee Y, Lee T W. Organic synapses for neuromorphic electronics: from brain-inspired computing to sensorimotor nervetronics. Accounts of Chemical Research, 2019, 52(4): 964−974

[161] Kim D W, Yang J C, Lee S, Park S. Neuromorphic processing of pressure signal using integrated sensor-synaptic device capable of selective and reversible short-and long-term plasticity operation. ACS Applied Materials & Interfaces, 2020, 12(20): 23207−23216

[162] Lee Y, Oh J Y, Lee T W. Neuromorphic skin based on emerging artificial synapses. Advanced Materials Technologies, 2022, 7(12): 2200193

[163] Jiménez-Fernández A, Cerezuela-Escudero E, Miró-Amarante L,

Domínguez-Morales M J, de Asís Gómez-Rodríguez F, Linares-Barranco A, Jiménez-Moreno G. A binaural neuromorphic auditory sensor for FPGA: a spike signal processing approach. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(4): 804−818

[164] Yamada R, Kuba H. Dendritic synapse geometry optimizes binaural computation in a sound localization circuit. Science Advances, 2021, 7(48): eabh0024

[165] Dalmas E, Danneville F, Bocquet M, Loyez C. Neuromorphic coincidence detector for interaural time difference encoding and sound DOA estimation. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 6505211

[166] Liu Y, Li E, Wang X, Chen Q, Zhou Y, Hu Y, Chen G, Chen H, Guo T. Self-powered artificial auditory pathway for intelligent neuromorphic computing and sound detection. Nano Energy, 2020, 78: 105403

[167] Duan G, Huang S, Feng Z, Xie P, Zhang F, Zhou Y, Han S T. Three-terminal artificial olfactory sensors based on emerging materials: mechanism and application. Advanced Functional Materials, 2023, 33(10): 202209969

[168] Song H W, Moon D, Won Y, Cha Y K, Yoo J, Park T H, Oh J H. A pattern recognition artificial olfactory system based on human olfactory receptors and organic synaptic devices. Science Advances, 2024, 10(21): eadl2882

[169] Podrażka M, Bączyńska E, Kundys M, Jeleń P S, Witkowska Nery E. Electronic tongue—a tool for all tastes? Biosensors, 2018, 8(1): 3

[170] Han J K, Park S C, Yu J M, Ahn J H, Choi Y K. A bioinspired artificial gustatory neuron for a neuromorphic based electronic tongue. Nano Letters, 2022, 22(13): 5244−5251

[171] Mayr C, Höppner S, Furber S. SpiNNaker 2: a 10 million core processor system for brain simulation and machine learning-keynote presentation. In: Pedersen J B, Chalmers K, Broenink J F, Vinter B, Vella K, Welch P H, Smith M L, Skovhede K, eds. Communicating Process Architectures 2017 & 2018. Washington: IOS Press, 2019: 277−280

[172] Davies M, Srinivasa N, Lin T H, Chinya G, Cao Y, et al. Loihi: a neuromorphic manycore processor with on-chip learning. IEEE Micro, 2018, 38(1): 82−99

[173] Merolla P A, Arthur J V, Alvarez-Icaza R, Cassidy A S, Sawada J, Akopyan F, Jackson B L, Imam N, Guo C, Nakamura Y, Brezzo B, Vo I, Esser S K, Appuswamy R, Taba B, Amir A, Flickner M D, Risk W P, Manohar R, Modha D S. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345(6197): 668−673

[174] Schoepe T, Gutierrez-Galan D, Dominguez-Morales J P, Greatorex H, Jimenez-Fernandez A, Linares-Barranco A, Chicca E. Closed-loop sound source localization in neuromorphic systems. Neuromorphic Computing and Engineering, 2023, 3(2): 024009

[175] Paredes-Vallés F, Hagenaars J J, Dupeyroux J, Stroobants S, Xu Y, de Croon G C H E. Fully neuromorphic vision and control for autonomous drone flight. Science Robotics, 2024, 9(90): eadi0591

[176] Gehrig D, Scaramuzza D. Low-latency automotive vision with event cameras. Nature, 2024, 629(8014): 1034−1040

[177] Ghosh S, D'Angelo G, Glover A, Iacono M, Niebur E, Bartolozzi C. Event-driven proto-object based saliency in 3d space to attract a robot's attention. Scientific Reports, 2022, 12(1): 7645

[178] Renner A, Evanusa M, Orchard G, Sandamirskaya Y. Event-based attention and tracking on neuromorphic hardware. In: Proceedings of the 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems. 2020

[179] McReynolds B, Graca R, Oliver R, Nishiguchi M, Delbruck T. Demystifying event-based sensor biasing to optimize signal to noise for space domain awareness. In: Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference. 2023

[180] Ziegler A, Joseph D, Gossard T, Moldovan E, Zell A. BiasBench: a reproducible benchmark for tuning the biases of event cameras. In: Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2025, 4994−5003

[181] Shoouri S, Taba M T, Kim H S. Adaptive LiDAR scanning: harnessing temporal cues for efficient 3D object detection via multi-modal fusion. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence. 2026

[182] Yang D, Liu Y, Chen Q, Chen M, Zhan S, Cheung N K, Chan H Y, Wang Z, Li W J. Development of the high angular resolution 360° LiDAR based on scanning MEMS mirror. Scientific Reports, 2023, 13(1): 1540

[183] D'Angelo G, Perrett A, Iacono M, Furber S, Bartolozzi C. Event driven bio-inspired attentive system for the iCub humanoid robot on spinnaker. Neuromorphic Computing and Engineering, 2022, 2(2): 024008

[184] Islam J, Wang R, Sattar J. SVAM: saliency-guided visual attention modeling by autonomous underwater robot. In: Proceedings of the 18th Robotics: Science and Systems. 2022

[185] Jin S, Dai X, Meng Q. "focusing on the right regions"—Guided saliency prediction for visual SLAM. Expert Systems with Applications, 2023, 213: 119068

[186] Ding N, Zhang C, Eskandarian A. SalienDet: a saliency-based feature enhancement algorithm for object detection for autonomous driving. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 2624−2635

[187] Gutierrez-Galan D, Bartolozzi C, Dominguez-Morales J P, Jimenez-Fernandez A, Linares-Barranco A. Towards the neuromorphic implementation of the auditory perception in the iCub robotic platform. In: Proceedings of the 2022 Annual Neuro-Inspired Computational Elements Conference. 2022, 11−12

[188] Gao W, Emaminejad S, Nyein H Y Y, Challa S, Chen K, Peck A, Fahad H M, Ota H, Shiraki H, Kiriya D, Lien D H, Brooks G A, Davis R W, Javey A. Fully integrated wearable sensor arrays for multiplexed *in situ* perspiration analysis. Nature, 2016, 529(7587): 509−514

[189] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 7132−7141

[190] Chen Y, Dai X, Liu M, Chen D, Yuan L, Liu Z. Dynamic convolution: attention over convolution kernels. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 11030−11039

[191] Yang Z, Mondal S, Ahn S, Xue R, Zelinsky G, Hoai M, Samaras D. Unifying top-down and bottom-up scanpath prediction using transformers. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 1683−1693

[192] Ilhan F, Su G, Tekin S F, Huang T, Hu S, Liu L. Resource-

efficient transformer pruning for finetuning of large models. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 16206−16215

[193] Ye X, Gan Y, Ge Y, Zhang X P, Tang Y. ATP-LLaVA: adaptive token pruning for large vision language models. In: Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025, 24972−24982

[194] Liu S, Guo B, Ma K, Yu Z, Du J. AdaSpring: context-adaptive and runtime-evolutionary deep model compression for mobile applications. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(1): 24

[195] Wolfe J M. Guided search 6.0: an updated model of visual search. Psychonomic Bulletin & Review, 2021, 28(4): 1060−1092

[196] Li Z. A saliency map in primary visual cortex. Trends in Cognitive Sciences, 2002, 6(1): 9−16

[197] Correa Á, Lupiáñez J, Madrid E, Tudela P. Temporal attention enhances early visual processing: a review and new evidence from event-related potentials. Brain Research, 2006, 1076(1): 116−128

[198] Parkhurst D, Law K, Niebur E. Modeling the role of salience in the allocation of overt visual attention. Vision Research, 2002, 42(1): 107−123

[199] Itti L, Koch C. Computational modelling of visual attention. Nature Reviews Neuroscience, 2001, 2(3): 194−203

[200] Rutishauser U, Walther D, Koch C, Perona P. Is bottom-up attention useful for object recognition? In: Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004

[201] Zhao J X, Liu J J, Fan D P, Cao Y, Yang J, Cheng M M. EGNet: edge guidance network for salient object detection. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. 2019, 8779−8788

[202] Cheng M M, Mitra N J, Huang X, Torr P H S, Hu S M. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 569−582

[203] Kaya E M, Elhilali M. Investigating bottom-up auditory attention. Frontiers in Human Neuroscience, 2014, 8: 327

[204] Chen S, Zhou Z, Hou K, Wu X, He Q, Tang C G, Li T, Zhang X, Jie J, Gao Z, Mathews N, Leong W L. Artificial organic afferent nerves enable closed-loop tactile feedback for intelligent robot. Nature Communications, 2024, 15(1): 7056

[205] Bartolozzi C, Indiveri G, Donati E. Embodied neuromorphic intelligence. Nature Communications, 2022, 13(1): 1024

[206] Wachowiak M. All in a sniff: olfaction as a model for active sensing. Neuron, 2011, 71(6): 962−973

[207] Aroudi A, Doclo S. Cognitive-driven binaural beamforming using EEG-based auditory attention decoding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 862−875

[208] Vanattou-Saïfoudine N, Han C, Krause R, Vasilaki E, von der Behrens W, Indiveri G. A robust model of stimulus-specific adaptation validated on neuromorphic hardware. Scientific Reports, 2021, 11(1): 17904

[209] Spratling M W. A review of predictive coding algorithms. Brain and Cognition, 2017, 112: 92−97

[210] Gabhart K M, Xiong Y, Bastos A M. Predictive coding: a more cognitive process than we thought?. Trends in Cognitive Sciences, 2025, 29(7): 627−640

[211] McDonald J J, Green J J, Störmer V S, Hillyard S A. Cross-modal spatial cueing of attention influences visual perception. In: Murray M M, Wallace M T, eds. The Neural Bases of Multisensory Processes. Boca Raton: CRC Press, 2012

[212] Shi K, He S, Shi Z, Chen A, Xiong Z, Chen J, Luo J. Radar and camera fusion for object detection and tracking: a comprehensive survey. IEEE Communications Surveys & Tutorials, 2026, 28: 3478−3520

[213] Lai H, Yin P, Scherer S. AdaFusion: visual-LiDAR fusion with adaptive weights for place recognition. IEEE Robotics and Automation Letters, 2022, 7(4): 12038−12045

[214] Vora S, Lang A H, Helou B, Beijbom O. PointPainting: sequential fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 4604−4612

[215] Afouras T, Chung J S, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8717−8727

[216] Xu P, Zhu X, Clifton D A. Multimodal learning with transformers: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113−12132

[217] Zong Z, Ma B, Shen D, Song G, Shao H, Jiang D, Li H, Liu Y. MoVA: adapting mixture of vision experts to multimodal context. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2024, 3282

[218] Xu W, Dong Z, Wang R, Xu X, Zhang Z. GateM^{2}Former: gated feature selection and expert modeling in multimodal emotion recognition. In: Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. 2025, 1−5

[219] Harlow K, Jang H, Barfoot T D, Kim A, Heckman C. A new wave in robotics: survey on recent MmWave radar applications in robotics. IEEE Transactions on Robotics, 2024, 40: 4544−4560

[220] Rashid N, Mortlock T, Al Faruque M A. Stress detection using context-aware sensor fusion from wearable devices. IEEE Internet of Things Journal, 2023, 10(16): 14114−14127

[221] Zhang Y, Carballo A, Yang H, Takeda K. Perception and sensing for autonomous vehicles under adverse weather conditions: a survey. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 196: 146−177

[222] Ambika N. Enhanced assistive technology on audio-visual speech recognition for the hearing impaired. In: Khang A, Shah V, Rani S, eds. Handbook of Research on AI-Based Technologies and Applications in the Era of the Metaverse. Hershey: IGI Global, 2023, 331−341

[223] Zhang Z, Zhou Z, Wang H, Zhang Z, Huang H, Cao Q. Grasp stability assessment through attention-guided cross-modality fusion and transfer learning. In: Proceedings of 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2023, 9472−9479

[224] Tayarani-Najaran M H, Schmuker M. Event-based sensing and signal processing in the visual, auditory, and olfactory domain: a review. Frontiers in Neural Circuits, 2021, 15: 610446

[225] Sandamirskaya Y, Kaboli M, Conradt J, Celikel T. Neuromorphic computing hardware and neural architectures for robotics. Science Robotics, 2022, 7(67): eabl8419

[226] Lakshmi A, Chakraborty A, Thakur C S. Neuromorphic vision: from sensors to event-based algorithms. WIREs Data Mining and

Knowledge Discovery, 2019, 9(4): e1310
[227] Wang T, Jin T, Lin W, Lin Y, Liu H, Yue T, Tian Y, Li L, Zhang Q, Lee C. Multimodal sensors enabled autonomous soft robotic system with self-adaptive manipulation. ACS Nano, 2024, 18(14): 9980−9996
[228] Wang H, Sun B, Ge S S, Su J, Jin M L. On non-von Neumann flexible neuromorphic vision sensors. npj Flexible Electronics, 2024, 8(1): 28
[229] Osswald M, Ieng S H, Benosman R, Indiveri G. A spiking neural network model of 3D perception for event-based neuromorphic stereo vision systems. Scientific Reports, 2017, 7(1): 40703
[230] Bartolozzi C, Indiveri G. A neuromorphic selective attention architecture with dynamic synapses and integrate-and-fire neurons. In: Proceedings of the Brain Inspired Cognitive Systems. 2004, 1−6
[231] Paul C. Morphological computation: a basis for the analysis of morphology and control requirements. Robotics and Autonomous Systems, 2006, 54(8): 619−630
[232] Sun J, Lerner E, Tighe B, Middlemist C, Zhao J. Embedded shape morphing for morphologically adaptive robots. Nature Communications, 2023, 14(1): 6023
[233] Pawlak W A, Howard N. Neuromorphic algorithms for brain implants: a review. Frontiers in Neuroscience, 2025, 19: 1570104
[234] Buschman T J, Kastner S. From behavior to neural dynamics: an integrated theory of attention. Neuron, 2015, 88(1): 127−144
[235] Abel R, Ullman S. Biologically inspired learning model for instructed vision. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2024, 1440
[236] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2014
[237] Knudsen E I. Fundamental components of attention. Annual Review of Neuroscience, 2007, 30: 57−78
[238] Braver T S, Reynolds J R, Donaldson D I. Neural mechanisms of transient and sustained cognitive control during task switching. Neuron, 2003, 39(4): 713−726
[239] Lindsay G W. Attention in psychology, neuroscience, and machine learning. Frontiers in Computational Neuroscience, 2020, 14: 29
[240] Wen H, Li Y, Zhang Z, Jiang S, Ye X, Ouyang Y, Zhang Y, Liu Y. AdaptiveNet: post-deployment neural architecture adaptation for diverse edge environments. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. 2023, 28
[241] Rao Y, Liu Z, Zhao W, Zhou J, Lu J. Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10883−10897
[242] Ma K, Tang J, Guo B, Dang F, Liu S, Zhu Z, Wu L, Fang C, Chen Y C, Yu Z, Liu Y. Surgeon: memory-adaptive fully test-time adaptation via dynamic activation sparsity. In: Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025, 30514−30523
[243] Qin Y, Zhou E, Liu Q, Yin Z, Sheng L, Zhang R, Qiao Y, Shao J. MP5: a multi-modal open-ended embodied system in minecraft via active perception. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, 16307−16316
[244] Hassani S, Dackermann U. A systematic review of optimization algorithms for structural health monitoring and optimal sensor placement. Sensors, 2023, 23(6): 3293
[245] Liu Y, Liu L, Zheng Y, Liu Y, Dang F, Li N, Ma K. Embodied navigation. Science China Information Sciences, 2025, 68(4): 141101
[246] Kurz A, Neff T, Lv Z, Zollhöfer M, Steinberger M. AdaNeRF: adaptive sampling for real-time rendering of neural radiance fields. In: Proceedings of the 17th European Conference on Computer Vision. 2022, 254−270
[247] Pfeifer R, Scheier C. Sensory—motor coordination: the metaphor and beyond. Robotics and Autonomous Systems, 1997, 20(2-4): 157−178
[248] Lagarde J, Kelso J A S. Binding of movement, sound and touch: multimodal coordination dynamics. Experimental Brain Research, 2006, 173(4): 673−688
[249] Farkaš I, Malík T, Rebrová K. Grounding the meanings in sensorimotor behavior using reinforcement learning. Frontiers in Neurorobotics, 2012, 6: 1
[250] Naros G, Naros I, Grimm F, Ziemann U, Gharabaghi A. Reinforcement learning of self-regulated sensorimotor β-oscillations improves motor performance. Neuroimage, 2016, 134: 142−152
[251] Lazareva O F, Shimizu T, Wasserman E A. How Animals See the World: Comparative Behavior, Biology, and Evolution of Vision. Oxford: Oxford University Press, 2012
[252] Wu L, Guo B, Zhang Q, Sun Z, Zhang J, Yu Z. Learning to self-reconfigure for freeform modular robots via altruism proximal policy optimization. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. 2023, 610
[253] Wu L, Guo B, Xu R, Liu S, Ding Y, Dang Z. Emergence of crowd modular robotics: a ubiquitous computing perspective. Scientia Sinica Informationis, 2023, 53(11): 2107−2151
[254] Hou L, Gao L, Wu Y, Chang Y. A survey on evaluation of embodied AI. Authorea, 2026
[255] Smagulova K, Elsheikh A, Silva D A, Fouda M E, Eltawil A M. Efficient and real-time perception: a survey on end-to-end event-based object detection in autonomous driving. Frontiers in Robotics and AI, 2025, 12: 1674421



Ruonan XU received the Bachelor's degree in computer science and technology from Northwestern Polytechnical University, China. She is currently working toward the PhD degree in computer science at Northwestern Polytechnical University, China. Her research interests include crowdsensing, deep reinforcement learning, and embodied intelligence.



Bin GUO received the PhD degree in computer science from Keio University, Japan in 2009, and then was a postdoc researcher with Institut Telecom SudParis, France. He is a professor with Northwestern Polytechnical University, China. His research interests include AIoT, ubiquitous computing, mobile crowd sensing, and HCI. He has served as an associate editor of the IEEE Communications Magazine and the IEEE Transactions on Human-Machine-Systems, the guest editor of the ACM Transactions on Intelligent Systems and Technology and the IEEE Internet of Things, the general co-chair of IEEE UIC'15, and the program chair of IEEE CPSCom'16, ANT'14, and UIC'13. He is a senior member of the IEEE.



Ke MA received the BE degree from Northwestern Polytechnical University, China in 2021. He is currently pursuing a PhD degree in computer science at Northwestern Polytechnical University, China. His research interests focus on deep learning for resource-constrained environments.



Lei WU received the Bachelor's degree in information and computing science from Northwestern Polytechnical University, China in 2021, where he is currently pursuing the PhD degree. His research interests include mobile crowdsensing, embodied intelligence, and reconfigurable robotics.



Yasan DING received his BE degree and PhD degree from Northwestern Polytechnical University, China in 2018 and 2024, respectively. His main research interests include crowd intelligence, and social data mining.



Yao JING received the Master's degree in computer science and technology from Northwestern Polytechnical University, China in 2018, and the PhD degree in computer science from the same university in 2025. Her research interests include urban computing, mobile crowdsensing, and reinforcement learning.



Zhiwen YU received the PhD degree in Engineering in computer science and technology from Northwestern Polytechnical University, China in 2005, where he is currently a professor. He was a post-doctoral researcher at the Information Technology Center, Nagoya University, Japan in 2006–2007, and a research fellow at the Academic Center for Computing and Media Studies, Kyoto University, Japan from February 2007 to January 2009. His research interests include pervasive computing, context-aware systems, human-computer interaction, mobile social networks, and personalization. He serves as associate editor or editorial board member of IEEE Communications Magazine, IEEE Transactions on Human-Machine Systems, Personal and Ubiquitous Computing, and Entertainment Computing. He is an IEEE Senior Member, ACM member, and CCF council member.