

Editorial

Open Access



Toward the next frontier of embodied AI

Rui Fan¹ , Mingjian Sun^{2,3,4,5,6}, George Giakos⁷

¹Department of Control Science & Engineering, Tongji University, Shanghai 201804, China.

²Harbin Institute of Technology (Weihai) Qingdao Innovation and Development Base, Qingdao 266109, Shandong, China.

³Department of Control Science and Engineering, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China.

⁴Harbin Institute of Technology, Weihai, Weihai 264200, Shandong, China.

⁵Harbin Institute of Technology Suzhou Research Institute, Suzhou 215000, Jiangsu, China.

⁶Shandong Laboratory of Advanced Biomaterials and Medical Devices in Weihai, Weihai 264209, Shandong, China.

⁷Department of Electrical & Computer Engineering, Manhattan University, University Park, PA 16802, USA.

Correspondence to: Prof. Rui Fan, Department of Control Science & Engineering, Tongji University, Shanghai 201804, China. E-mail: rui.fan@ieee.org

How to cite this article: Fan, R.; Sun, M.; Giakos, G. Toward the next frontier of embodied AI. *Intell. Robot.* 2025, 5, 859–63.
<https://dx.doi.org/10.20517/ir.2025.44>

Received: 4 Nov 2025 **Accepted:** 17 Nov 2025 **Published:** 2 Dec 2025

Academic Editor: Simon Yang **Copy Editor:** Ting-Ting Hu **Production Editor:** Ting-Ting Hu

Abstract

Embodied artificial intelligence has emerged as a transformative paradigm, marking a fundamental shift in artificial intelligence research toward systems that tightly couple perception, cognition, and action within real-world environments. This editorial emphasizes the growing significance of embodied artificial intelligence, introduces the key contributions presented in this Special Issue, and provides an overview of the current challenges and prospective research directions shaping the future of the field.

Keywords: Robotics, artificial intelligence, embodied AI

1. INTRODUCTION

Recently, embodied artificial intelligence has emerged as a transformative paradigm in artificial intelligence research, marking a significant shift toward systems that tightly couple perception, action, and interaction within physical environments^[1]. Progress in this field is fueled by interdisciplinary insights and state-of-the-art technologies, enabling the development of adaptive, autonomous agents that can navigate and learn from complex, real-world scenarios^[2]. This evolution signals a new chapter in artificial intelligence, wherein machines transcend passive observation to actively engage with their environments, exhibiting levels of situational understanding that increasingly resemble human cognitive processes^[3].



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



Taking autonomous driving as an example, embodied artificial intelligence enables vehicles not only to perceive their surroundings through advanced sensors such as LiDAR, Radar, and cameras, but also to dynamically adapt to evolving traffic conditions, interact safely with pedestrians, and respond effectively to unforeseen events^[4]. These intelligent agents leverage contextual understanding to make real-time decisions that prioritize both safety and operational efficiency^[5]. Similarly, service robots equipped with embodied intelligence can skillfully navigate domestic and public environments, comprehend natural language instructions, and engage in complex interactions with both objects and humans^[6]. Such advancements mark a significant step toward the meaningful integration of artificial intelligence into real-world environments.

This Special Issue aims to demonstrate recent advances in the rapidly evolving field of embodied artificial intelligence. We received eight submissions from researchers across the globe, reflecting the growing interest and momentum in this area. Following a rigorous peer-review process and valuable feedback from expert reviewers, four articles were selected for publication. These contributions primarily address key challenges in perception and localization, presenting novel insights and methodologies that advance the capabilities of intelligent embodied systems.

2. CONTRIBUTED ARTICLES

Among the four accepted articles, three deep learning-based studies focus on leveraging single-modal neural networks to perform salient object detection^[7], facial expression recognition^[8], and infrared object detection^[9], respectively. Although targeting different application domains, these studies share a unified vision: enabling machines to perceive, interpret, and respond to complex visual environments in real time. Each work introduces specialized network architectures designed to enhance perception accuracy while maintaining computational efficiency, reflecting a collective push toward deployable artificial intelligence in resource-constrained or dynamic conditions. A notable commonality among these studies lies in their emphasis on multi-scale feature learning. Whether through attention mechanisms, convolution-transformer fusion, or scale-adaptive modules, all three approaches integrate hierarchical visual cues to capture both global semantic context and fine-grained structural detail. This multi-scale strategy enables a crucial balance between semantic comprehension and spatial precision, an essential attribute for embodied artificial intelligence systems operating in unstructured or variable environments. Furthermore, each method demonstrates strong empirical performance on public benchmark datasets, underscoring the robustness and generalizability of the proposed designs.

Despite these shared foundations, each study makes distinct and complementary contributions. The study^[7] focuses on lightweight salient object detection, introducing scale-adaptive feature extraction and multi-scale feature aggregation modules to achieve an optimal trade-off between efficiency and accuracy. The study^[8] addresses facial expression recognition toward embodied artificial intelligence, proposing a multi-scale attention and convolution-transformer fusion network to enhance emotion-aware human-robot interaction^[10]. Finally, the study^[9] targets infrared object detection under adverse weather conditions, combining the MobileNetV3-YOLOv4 architecture with an image enhancement generative adversarial network to ensure high-precision detection on low-power edge devices.

Taken together, these studies collectively advance the frontier of efficient and adaptive visual perception. By addressing complementary aspects of perception, ranging from the semantic understanding of human emotions to the structural saliency of objects and the multi-modal robustness required under low-visibility conditions, they demonstrate a coherent progression toward unified, context-aware perception systems. Such efforts not only contribute to academic exploration but also carry significant practical implications for real-world applications, from intelligent vehicles and service robots to next-generation embodied agents capable of understanding and interacting with their surroundings in a human-like manner.

In addition, the remaining study^[11] explores the parallel implementation of a real-time visual simultaneous localization and mapping system through heterogeneous parallel computing. Although it does not rely on deep learning, the study shares several conceptual commonalities with the aforementioned works. First, it emphasizes computational efficiency and real-time performance, an objective aligned with the other contributions that enhance learning and perception efficiency through architectural optimization or multi-modal fusion. Second, similar to the deep learning-based studies, it contributes to the advancement of embodied intelligence and autonomous systems, where robust perception and low-latency computation are critical for deployment in complex real-world environments. Third, all four studies integrate vision with high-performance computing strategies to manage large-scale visual data: the deep learning approaches do so through large vision models and multi-modal learning, while this work adopts hardware-level parallelization. Overall, this study complements the learning-based approaches by addressing computational bottlenecks from a systems and hardware perspective, underscoring the importance of real-time, scalable visual processing for embodied artificial intelligence.

3. EXISTING CHALLENGES AND FUTURE RESEARCH TRENDS

Despite the remarkable progress in embodied artificial intelligence, several key challenges remain unresolved. First, most existing frameworks exhibit limited generalization and adaptability when deployed in diverse real-world environments. While deep learning-based methods have demonstrated impressive capabilities, they often depend heavily on large-scale annotated datasets and struggle with continual learning or domain transfer. This reliance frequently leads to catastrophic forgetting during long-term autonomous operation. Second, current multi-modal perception systems lack efficient mechanisms for cross-modal understanding and fusion, especially in dynamic scenarios where heterogeneous sensor data present misaligned spatio-temporal characteristics. Third, intelligent systems face significant constraints related to computational and energy efficiency. The widespread dependence on large-scale vision-language models and complex neural architectures hampers their real-time applicability on resource-constrained platforms, such as mobile robots, aerial drones, and intelligent vehicles. Finally, there remains a gap between algorithmic innovation and hardware-level optimization. Most research emphasizes model design while neglecting the computational parallelization, scheduling, and optimization strategies required for efficient deployment on heterogeneous devices.

Future research in embodied artificial intelligence is expected to evolve along several promising directions. First, lifelong and continuous learning will become a central objective, enabling embodied agents to incrementally acquire new skills and adapt to changing environments without retraining from scratch^[12]. Second, multi-modal and cross-domain integration will deepen, aiming to build unified representations that effectively combine vision, language, and sensory cues for robust reasoning and decision-making^[13-15]. Third, the development of large-scale foundation models specifically tailored for embodied perception, particularly those inspired by the human dual-stream visual processing hypothesis, will catalyze progress toward generalizable understanding and interaction across diverse tasks and contexts^[16]. Fourth, the design of lightweight and energy-efficient models will be essential for deploying complex visual-language architectures on embedded platforms, necessitating advances in compression, pruning, and knowledge distillation techniques^[17]. Fifth, hardware-software co-optimization is expected to receive growing attention, with the integration of parallel computing strategies, such as Compute Unified Device Architecture (CUDA)-based heterogeneous acceleration and graphics processing unit (GPU)/tensor processing unit (TPU) adaptation, playing a vital role in addressing the computational bottlenecks that currently constrain real-time embodied intelligence^[11]. Finally, future embodied artificial intelligence systems must prioritize interpretability and safety to ensure transparent decision-making and foster trustworthy interactions between intelligent agents and humans in complex, unstructured environments.

4. CONCLUSION

In summary, this editorial underscored the pressing need for advancing embodied artificial intelligence, highlighted the key contributions of the accepted articles in this special issue, and briefly discussed both the current challenges and promising future research directions in the field.

DECLARATIONS

Authors' contributions

Substantial contributions to conception and design of the editorial: Fan, R.; Sun, M.; Giakos, G.

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the Fundamental Research Funds for the Central Universities and the Xiaomi Young Talents Program.

Conflicts of interest

Fan, R. is Guest Editor of the Special Issue "Recent Advances in Embodied Artificial Intelligence" and Editorial Board Member of the journal *Intelligence & Robotics*. Giakos, G. is Guest Editor of the Special Issue "Recent Advances in Embodied Artificial Intelligence". Neither of them was involved in any steps of editorial processing, notably including manuscript handling and decision-making. The other authors have declared that they have no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Li, J.; Yang, S. X. Digital twins to embodied artificial intelligence: review and perspective. *Intell. Robot.* **2025**, 5, 202-27. DOI
2. Chen, W.; Chi, W.; Ji, S.; et al. A survey of autonomous robots and multi-robot navigation: perception, planning and collaboration. *Biomim. Intell. Robot.* **2025**, 5, 100203. DOI
3. Bin, T.; Yan, H.; Wang, N.; Nikolić, M. N.; Yao, J.; Zhang, T. A survey on the visual perception of humanoid robot. *Biomim. Intell. Robot.* **2025**, 5, 100197. DOI
4. Fan, R.; Guo, S.; Bocus, M. J. *Autonomous driving perception*. Cham, Switzerland: Springer, 2023. https://scholar.google.com/scholar?q=Autonomous+driving+perception.+Cham,+Switzerland:+Springer+2023&hl=zh-CN&as_sdt=0&as_vis=1&oi=scholar (accessed 2025-11-24).
5. Huang, Y.; Fan, D.; Duan, H.; et al. Human-like dexterous manipulation for anthropomorphic five-fingered hands: a review. *Biomim. Intell. Robot.* **2025**, 5, 100212. DOI
6. Mon-Williams, R.; Li, G.; Long, R.; Du, W.; Lucas, C. G. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nat. Mach. Intell.* **2025**, 7, 592-601. DOI PubMed PMC
7. Liu, Z.; Zhao, W.; Jia, N.; Liu, X.; Yang, J. SAnet: scale-adaptive network for lightweight salient object detection. *Intell. Robot.* **2024**, 4, 503-23. DOI

8. He, H.; Liao, R.; Li, Y. MSAFNet: a novel approach to facial expression recognition in embodied AI systems. *Intell. Robot.* **2025**, *5*, 313-32. [DOI](#)
9. Zhuang, T.; Liang, X.; Xue, B.; Tang, X. An in-vehicle real-time infrared object detection system based on deep learning with resource-constrained hardware. *Intell. Robot.* **2024**, *4*, 276-92. [DOI](#)
10. Zhang, C.; Chen, J.; Li, J.; Peng, Y.; Mao, Z. Large language models for human-robot interaction: a review. *Biomim. Intell. Robot.* **2023**, *3*, 100131. [DOI](#)
11. Liu, H.; Dong, Y.; Hou, C.; et al. Parallel implementation for real-time visual SLAM systems based on heterogeneous computing. *Intell. Robot.* **2024**, *4*, 256-75. [DOI](#)
12. Chen, Q.; Wang, C.; Wang, D.; Zhang, T.; Li, W.; He, X. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. *arXiv* **2004**, arXiv:2411.15432. [DOI](#)
13. Huang, J.; Li, J.; Jia, N.; et al. RoadFormer+: delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion. *IEEE. Trans. Intell. Veh.* **2025**, *10*, 3156-65. [DOI](#)
14. Liu, Y.; Chen, Q.; Albanie, S. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021. pp 14954-64. https://openaccess.thecvf.com/content/CVPR2021/papers/Liu_Adaptive_Cross (accessed 2025-11-24).
15. Guo, S.; Long, Z.; Wu, Z.; Chen, Q.; Pitas, I.; Fan, R. LIX: implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving. *IEEE. Trans. Image. Process.* **2025**, *34*, 7250-63. [DOI](#) [PubMed](#)
16. Liu, C.; Chen, Q.; Fan, R. Playing to vision foundation model's strengths in stereo matching. *IEEE*, 2024; pp 1-12. [DOI](#)
17. Liu, H.; Galindo, M.; Xie, H.; et al. Lightweight deep learning for resource-constrained environments: a survey. *ACM. Comput. Surv.* **2024**, *56*, 1-42. [DOI](#)