VLAS: From Individual to Social Intelligence via Social Radar-Augmented VLA Agents

1st Lili Fan
School of Artificial Intelligence
Beijing Institute of Technology
Beijing, China
lilifan@bit.edu.cn

2nd Zeming Jin
School of Artificial Intelligence
Beijing Institute of Technology
Beijing, China
3220251191@bit.edu.cn

3rd Liyuan Fan
Ship Electronic Engineering Technology
China Ship Research Academy
Beijing, China
fanliyuanfly0927@163.com

4th Wei Wang*
School of Languages and Communication Studies
Beijing Jiaotong University
Beijing, China
weiwang3@bjtu.edu.cn

5th Levente Kovacs
John von Neumann Faculty of Informatics
Obuda University
Budapest, Hungary
kovacs@uni-obuda.hu

***Abstract*—Recent advancements in Vision–Language–Action (VLA) models have enabled embodied agents to perceive visual scenes, follow natural language commands, and perform low-level control in an end-to-end fashion. However, existing VLA agents primarily operate as individual intelligence systems, lacking awareness of collective human behavior, social context, and group-level intent. In this work, we propose VLAS (Vision–Language–Action–Social), a new paradigm that incorporates Social Radar—a mechanism that captures signals from social media, crowdsourced distress calls, emotional sentiment, and group dynamics—into the core VLA pipeline. By fusing social cues with multimodal sensor data, VLAS agents gain the ability to understand collective intent, infer emotional states, and engage in socially aligned decision-making. This integration bridges the gap between individual embodiment and social intelligence, closing the loop across four key dimensions: language, perception, reasoning, and interaction. We detail the architectural extensions required to support Social Radar integration, survey representative application scenarios (e.g., disaster response, urban safety, battlefield cognition), and outline the challenges and opportunities of building socially intelligent embodied agents.**

***Index Terms*—social intelligence, multimodal perception, social radar, Vision–Language–Action (VLA), embodied agents.**

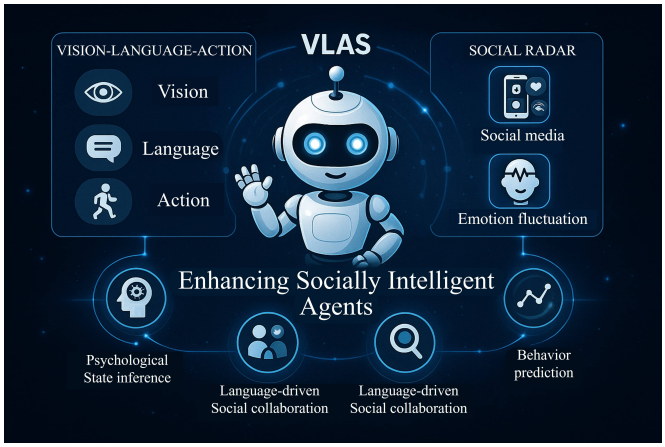

Fig. 1. Schematic diagram of the VLAS (Vision–Language–Action–Social) formed by the integration of social radar and VLA.

I. Introduction

In recent years, VLA models have achieved remarkable progress in the fields of embodied intelligence and autonomous systems [1], enabling agents to integrate visual perception, language understanding, and action decision-making within a unified framework to support end-to-end execution of complex tasks. However, despite their strong generalization and instruction-following capabilities in simulated environments, VLA systems still face several critical challenges in adapting to real-world, complex scenarios and engaging in social collaboration. Firstly, the language input of current VLA systems largely depends on manually defined or static navigation instructions, making it difficult to perceive and respond to emergent events, contextual changes, or group-level behavioral intentions in real time. Secondly, their sensing modalities primarily focus on camera, lidar and millimeter-wave radar [2], which, while effective at capturing the physical structure and dynamics of the environment, lack the capacity to model high-level semantics such as public opinion trends, social emotions, and collective behavioral patterns. On the reasoning side, although recent studies have introduced mechanisms like Chain-of-Thought to enhance interpretability [3], current systems still struggle to handle complex decision-making scenarios involving social psychology, emotional fluctuations, and cross-agent coordination. Furthermore, most existing agents rely on static command-response interactions or simplified V2V cooperation protocols [4], and

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB3209801 and by the National Natural Science Foundation of China under Grant 52402490.

are incapable of perceiving or adapting to feedback signals and dynamic changes in real social interactions.

To overcome these limitations, this paper proposes the integration of a Social Radar mechanism into VLA systems to construct a new generation of agents equipped with social perception and collaboration capabilities [5]. Social radar actively gathers signals from social media, real-time semantic help requests, and group-level emotional fluctuations, thereby significantly enhancing the timeliness and contextual relevance of language inputs. At the perception level, it enables behavior prediction and public opinion recognition, extending the agent's cognitive boundary from the physical world to the social context. At the reasoning level, social signals provide a data foundation for modeling group decision logic and psychological state inference, equipping agents to understand and respond to social behaviors. In terms of interaction, social radar promotes a shift from command-driven responses to language-driven social collaboration, allowing agents to adaptively respond to dynamic social feedback and engage in context-aware coordination with other agents.

In summary, the introduction of social radar not only provides VLA systems with a new modality of perception and a novel cognitive path, but also drives a paradigm shift from individual intelligence to socially intelligent agents. The architecture of VLAS is shown in Fig. 1. This paper systematically presents the design principles, key capabilities, and representative application scenarios of the proposed social radar-augmented VLA architecture, and explores its potential value and future challenges in socially complex environments such as disaster response, urban safety, and cognitive warfare.

II. Related Work

A. VLA Systems: Architectures and Evolution

Multimodal foundation models [6] have become a core approach in VLA learning, driving the integration and collaboration of vision, language and action. With the rise of multimodal foundation models like CLIP [7], Flamingo [8], and InternVL [9], VLA systems have advanced rapidly across domains such as autonomous driving, robotics, and embodied AI. Current research in VLA generally follows three paradigms [10]: (1) Modular architectures, which decompose perception, language, and control into discrete stages; (2) End-to-end systems, which fuse multimodal inputs within unified Transformer frameworks to map perception directly to control; and (3) Reasoning-centric agents, which incorporate Chain-of-Thought and memory for causal reasoning and interpretability. Despite progress in modality fusion and language interaction, these systems remain limited in their ability to perceive social context, collective behavior, and emotional dynamics within socially complex environments.

B. Social Radar: Model and Applications

"Social Radar" is an emerging sensing paradigm that integrates social science and information technology to enable the perception and reasoning of multi-level social signals such as cognition, emotion, behavior, and intent. Initially proposed by Schramm [5] and later expanded within the CPSS framework [11], it collects individual and group-level information from sources like social media, news, and online forums. Using techniques such as keyword expansion, sentiment analysis, and trend modeling, it builds a soft sensing infrastructure suited to cyberspace dynamics. Demonstrating real-time insight and predictive capabilities, social radar has been applied in political decision-making, urban governance, transportation, and disaster response. As a complement to physical sensing, it provides AI systems with a secondary channel for accessing social cognition [12].

Although VLA systems have advanced in multimodal fusion and control, their reliance on physical-domain perception limits understanding of higher-level social constructs such as intent, semantics, and emotional dynamics [13]. They struggle to process unstructured social signals, model collective intentions, or respond to dynamic social feedback. Existing interaction paradigms remain reactive and task-driven, limiting use in complex social contexts. To address these limitations, we propose VLAS—an enhanced VLA architecture augmented with social radar—to enable the transition from individual to socially intelligent agents.

III. System Framework of VLAS

To address the limitations of VLA systems in real-world social environments, this paper proposes VLAS—a Social Radar-augmented VLA architecture that equips agents with social perception, contextual reasoning, and collaborative action capabilities. Building on the traditional vision, language, and action pipeline, VLAS introduces Social Radar as a fourth modality to capture emotional dynamics, collective behavior, and real-time social signals. By integrating these signals into all core modules, VLAS enables a shift from individual task execution to social scene understanding and coordination. Similar to the optimization strategies used in remote sensing image retrieval [14], the VLAS method also improves the system's ability to process complex data and enhances the synergy of multimodal inputs, enabling it to better cope with dynamic and complex social environments. This enhances agents' cognitive scope and decision-making for applications such as disaster response, urban safety, and human–machine interaction. Fig. 2 compares VLA module inputs and outputs before and after social radar integration, followed by a detailed analysis of the resulting functional and capability enhancements.

A. Vision Module: From Physical Perception to Social Situation Awareness

In the VLAS architecture, the vision module expands the conventional physical environment modeling syste into a dual-domain perception framework that integrates both physical and social dynamics. Beyond recognizing structured elements such as road geometry and traffic participants, the module incorporates heterogeneous inputs from Social Radar, including short videos from social media, crowd movement trajectories from urban surveillance systems, and geotagged

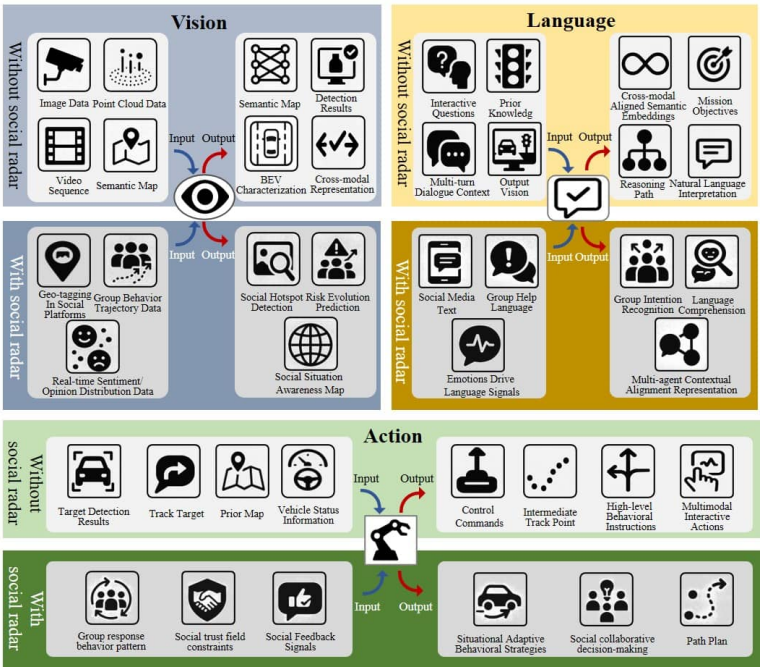

Fig. 2. The comparative diagram of inputs and outputs for the vision, language, and action modules in VLA before and after the integration of social radar.

images of emergent events. These multimodal sources enable the system to perceive spatially distributed social phenomena such as group behaviors, crowd aggregations, and emotional fluctuations. This module emphasizes the extraction of spatially localizable and temporally trackable social signals, including crowd density estimation, collective movement trend analysis, and detection of anomalous regional behaviors. By leveraging cross-modal spatial–temporal alignment mechanisms, it fuses physical imagery with socially derived data to construct dynamic social situation awareness maps. Furthermore, it generates structured "Human–Crowd–Scene" interaction graphs, which serve as spatial priors for downstream language understanding and behavioral decision-making processes. Compared to conventional visual systems that focus solely on static geometric structures, this module empowers the agent to anticipate social disturbances, predict shifts in collective behavior, and identify potential event triggers. These capabilities substantially enhance the system's adaptability and responsiveness in complex environments such as disaster response, urban security, and social collaboration.

B. Language Module: From Command Parsing to Social Contextualization

The language module in VLAS moves beyond the static task-oriented input paradigm of traditional VLA systems and is restructured into a socially contextualized language processing pipeline that supports real-time reasoning of multi-source inputs. It is capable of processing diversified language signals including social media texts, public voice channels, and crowd communication streams. The system is designed to handle unstructured, emotionally expressive, and multi-agent collaborative language forms. Through multi-turn dialogue modeling and cross-modal semantic alignment, the language module can infer collective intentions, emotional states, and cultural context embedded in natural language, and encode them into a semantic space suitable for downstream reasoning. To enhance the interpretable representation of collective intentions, emotions and contexts, the language module focuses on the identification and representation of relevant variables, while considering the role of Sociolinguistic Radar in defining the relationship between socially significant linguistic variants and social meanings [15]. Furthermore, the module supports role recognition and multi-agent language coordination, enabling contextual modulation, intention alignment, and semantic mapping in open-task scenarios. Ultimately, it provides critical inputs to both reasoning and behavior modules. It also serves as the primary interface for social dialogue, marking a functional shift from a command parser to a socially driven contextual hub.

C. Action Module: From Individual Decisions to Socially Coordinated Execution

In VLAS, the action module no longer relies solely on local state assessments and goal-driven path optimization. Instead, it is built upon a social feedback-driven decision-making mechanism that incorporates multi-agent coordination and ethical constraint modeling. Taking as input the social perception graphs from the vision module and collective intentions from the language module, along with behavioral history, social trust maps, and real-time feedback signals (e.g., requests or warnings) provided by Social Radar, the module generates behavior strategies that are contextually

Fig. 3. (a) Multi-agent search and rescue system in disaster response. (b) Multimodal Emergency Response Agent System for Urban Contingencies.

adaptive. The system can detect critical social states such as evacuation trends, unrest zones, and non-verbal expressions, and dynamically adjust behavior strategies including route replanning, speed modulation, and priority reassignment. It also supports collaborative execution among agents by negotiating intentions and coordinating responses via both verbal and non-verbal interaction protocols, thereby enhancing group-level consistency.,Ultimately, the action module does not output static control commands, but generates compound strategies that are socially acceptable, ethically aligned, and cooperation-oriented. This represents a paradigm shift from isolated controllers to embedded social participants, enabling safe, stable, and socially collaborative behavior in complex high-dynamic environments.

D. Social Radar Mechanism: From Data Processing to Semantic Modeling and Privacy Protection

At the data processing layer, the Social Radar performs preprocessing and feature extraction on heterogeneous data streams, mapping unstructured social signals into cross-modal aligned feature representations [16]. Textual data undergo noise filtering and anomaly detection before being fed into pretrained language models for robust semantic vectorization, while image and video data are cleaned and processed through vision models to extract visual features. For multi-source data, the system applies spatio-temporal alignment and correlation modeling to map social signals from social media dynamics, user help requests, and collective interaction logs into a unified feature space, and employs feature selection and dimensionality reduction techniques to reduce redundancy. Through this pipeline, social signals can be delivered to downstream modules in a normalized and stable form.

At the semantic modeling layer, the Social Radar integrates cross-modal embedding and semantic graph construction methods to transform linguistic content, emotional features, and collective behavioral patterns into structured representations of social context [17]. Textual information is semantically encoded using pretrained language models, while image and video data are processed by vision–language models to extract cross-modal features, and emotional cues are captured by Transformer-based emotion recognition models. These multimodal features, together with the outputs of intention inference and topic evolution modeling, are then integrated into a semantic graph, producing social semantic graphs that capture associations among events, emotions, and collective behaviors, while supporting causal reasoning and situational understanding. This representation provides contextual constraints and prior knowledge for the agent's language understanding and reasoning modules.

At the privacy protection layer, the Social Radar incorporates mechanisms such as differential privacy, federated learning, and hierarchical access control during data collection and modeling to reduce intrusiveness into individual privacy. At the feature extraction stage, the system introduces randomized noise into vectorized representations or applies anonymization strategies to weaken the traceability between user data and model outputs, thereby preventing the disclosure of identities and sensitive attributes. In addition, the system leverages a federated learning framework to perform model updates locally, thus reducing the risks of centralized data aggregation. While fulfilling privacy protection requirements and regulatory compliance, the system retains effective modeling capacity for social signals, ensuring that VLAS maintains efficient perception and reasoning performance.

In summary, the VLAS architecture integrates the Social Radar mechanism into the vision, language, and action modules to construct a closed-loop system for perception, interpretation, and decision-making in complex social environments. This framework not only equips intelligent agents with the ability to parse unstructured social contexts, but also enhances their responsiveness to emergent events and their capacity for generating socially coordinated behaviors. VLAS thus facilitates the paradigm shift from tool-based agents to socially intelligent agents. As a forward-looking architecture for intelligent systems, it lays the technological foundation for embodied agents with social perception capabilities and provides a unified modeling paradigm and engineering support for multimodal intelligence research under uncertainty

and interactional complexity.

IV. Application Scenarios

The VLAS framework demonstrates strong applicability in complex, socially embedded environments where physical sensing is insufficient. By integrating Social Radar with conventional VLA systems, agents acquire the capability to perceive human intent, emotional states, and public discourse in real time, enabling context-aware decision-making. This section presents two representative scenarios to illustrate how social signals can shape language inputs, inform high-level planning, and enhance behavior generation under dynamic and uncertain conditions, as shown in Fig. 3.

A. Emergency Search and Rescue

In post-disaster scenarios such as earthquakes and floods, the vision module integrates infrared imaging, millimeter-wave radar, and UAV video to construct real-time spatial representations of affected area, VLAS enhances rescue operations by incorporating a social radar function, providing timely and contextually relevant information, which further improves decision support in post-disaster environments [18]. By incorporating social radar, it fuses geotagged images and live social media feeds to semantically anchor distress signals to physical locations, generating a socially enhanced bird's-eye view map that informs rescue planning and language grounding. The language module processes distress signals from platforms such as Weibo and voice messages, extracting spatial cues, survival status, emotional states, and intent. Through emotion and intent inference, it supports real-time emergency assessment and multi-agent coordination, while also providing empathetic responses to foster human–AI trust. The action module generates a socially informed task graph based on fused inputs, prioritizing entrapment zones by assessing risk, resource constraints, and behavioral patterns. It enables coordinated deployment of robotic agents and adapts strategies through evacuation modeling and intent prediction, supporting rapid and resilient decision-making in dynamic disaster environments.

B. Emergency Response System for Urban Contingencies

In urban environments, large-scale public events such as viral gatherings or lantern fairs often trigger sudden crowd surges with concentrated spatial distribution and heightened emotional volatility. Incorporating social radar, the system accesses trending comments, geotagged posts, and event-related hashtags to identify anomalous attention hotspots. By fusing physical and social data, it constructs a multimodal alignment map linking spatial layout, online discourse, and collective behavior, providing semantic priors for downstream understanding and control. The language module extracts crowd sentiment and behavioral trends from social media streams using emotion classification and intent recognition models. It also generates real-time responses through public broadcasts and voice terminals, enhancing public trust and communication. Based on visual and linguistic inputs, the action module generates coordinated multi-agent strategies for urban safety. In areas with heightened social feedback or emotional volatility, it deploys robotic agents and broadcast units to implement interventions such as rerouting and flow control. Supported by behavior prediction, the system adapts strategies in real time, enabling resilient and socially aware response through a closed perception–action loop.

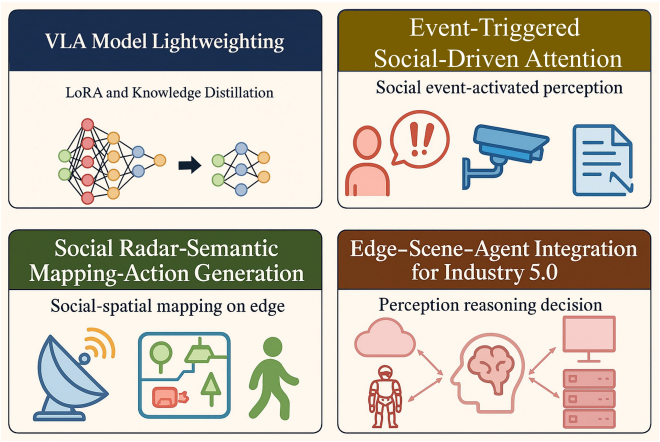

Fig. 4. Key directions for the future evolution of VLAS.

V. Future Direction

The proposed VLAS architecture provides a unified framework for integrating physical and social perception in high-level reasoning and coordination. To support its deployment in real-world, resource-constrained settings, especially in time-sensitive scenarios such as disaster relief, urban security, and mobile operations, future research should focus on the development of embedded, low-latency, edge-level intelligent

agents capable of performing vision–language–action inference and social signal fusion locally [19], as shown in Fig. 4.

1) Lightweight VLA via LoRA and Knowledge Distillation: Deploying VLAS on edge devices requires reducing model size and computational overhead while maintaining multimodal alignment and decision accuracy. Techniques such as Low-Rank Adaptation and knowledge distillation compress vision-language models without compromising cross-modal reasoning. Domain-specific fine-tuning on social–physical datasets further improves performance under memory and power constraints, supporting real-time operation on platforms such as mobile robots and emergency terminals.

2) Event-Triggered Perception and Social-Driven Attention: Unlike conventional always-on systems, edge agents in VLAS should adopt event-triggered processing guided by salient social radar cues. Emotional keywords, bursts of distress signals, or shifts in collective intent can activate targeted perception modules such as UAV patrols or semantic parsing, enabling energy-efficient yet responsive operation in dynamic physical–social environments.

3) Integrated Social Perception and Action Planning: Future VLAS systems should integrate social radar sensing, semantic mapping, and language-grounded action into a unified edge deployment pipeline. By enabling local reasoning over shared social–spatial representations, agents can interpret human intent, update maps in real time, and execute coordinated actions. This supports autonomous social–physical interaction loops in applications such as autonomous driving, smart security, and human–robot collaboration.

4) Toward Edge–Scene–Agent Integration for Industry 5.0: VLAS offers a foundation for a Scene–Agent–Edge–Cloud continuum, enabling embedded agents to sense, reason, and act locally while aligning with global intent through cloud coordination. This vision supports the principles of Industry 5.0 [20], [21], emphasizing real-time sensing, human-centric cognition, and context-aware autonomy. Its realization depends on advances in edge intelligence, low-power AI hardware, trust-aware collaboration, and scalable social perception.

VI. Conclusion

This paper introduces VLAS, a Vision–Language–Action–Social architecture that integrates social radar into conventional VLA systems. By incorporating unstructured inputs from social media, emotional states, and collective behaviors, VLAS expands agent perception beyond the physical domain to support reasoning and adaptive response in human-centered environments. Through enhanced language input, multimodal perception, and group-level inference, VLAS establishes a closed-loop cognitive framework for context-aware planning and collaborative decision-making. Applications in disaster response, urban risk management, and battlefield cognition illustrate its capacity to translate social signals into actionable insight.

Furthermore, this work outlines a forward-looking deployment pathway toward lightweight, event-triggered, and semantically coupled edge agents, positioning VLAS as a foundation for next-generation embedded AI. The proposed edge–scene–agent integration framework aligns with the vision of Industry 5.0, highlighting the critical role of social perception and multimodal fusion in enabling human-aligned, autonomous, and collaborative AI systems operating effectively in complex real-world environments.

References

[1] R. Sapkota *et al.*, "Vision-language-action models: concepts, progress, applications and challenges," *arXiv preprint arXiv:2505.04769*, 2025.

[2] L. Fan *et al.*, "4d mmwave radar for autonomous driving perception: A comprehensive survey," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4606–4620, 2024.

[3] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[4] J. Grosset *et al.*, "Multi-agent simulation of autonomous industrial vehicle fleets: towards dynamic task allocation in v2x cooperation mode," *Integrated Computer-Aided Engineering*, vol. 31, no. 3, pp. 249–266, 2024.

[5] W. L. Schramm and W. E. Porter, *Men, women, messages, and media : understanding human communication*, 2nd ed. Harper Row, 1982.

[6] L. Fan *et al.*, "Multimodal perception and decision-making systems for complex roads based on foundation models," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 11, pp. 6561–6569, 2024.

[7] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Jul. 2021, pp. 8748–8763.

[8] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[9] Z. Chen *et al.*, "Internvl: scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 185–24 198.

[10] S. Jiang *et al.*, "A survey on vision-language-action models for autonomous driving," *arXiv preprint arXiv:2506.24044*, 2025.

[11] J. J. Zhang *et al.*, "Cyber-physical-social systems: the state of the art and perspectives," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 829–840, 2018.

[12] L. Fan *et al.*, "Social radars: finding targets in cyberspace for cybersecurity," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 279–282, 2024.

[13] Y. Ma *et al.*, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2025.

[14] L. Fan, H. Zhao, and H. Zhao, "Global optimization: combining local loss with result ranking loss in remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 7011–7026, 2020.

[15] W. Wang *et al.*, "Sociolinguistic radar of phonological variation and social meaning: variables, quantitative methods, and prospects," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 6, pp. 7734–7741, 2024.

[16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[17] Y. Chen *et al.*, "A survey on multimodal knowledge graphs: Construction, completion and applications," *Mathematics*, vol. 11, no. 8, 2023.

[18] L. Fan *et al.*, "Social radars for social vision of intelligent vehicles: a new direction for vehicle research and development," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 3, pp. 4244–4248, 2024.

[19] X. Wang *et al.*, "Steps toward industry 5.0: Building "6s" parallel industries with cyber-physical-social intelligence," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 8, pp. 1692–1703, 2023.

[20] J. Xu *et al.*, "When embodied ai meets industry 5.0: human-centered smart manufacturing," *IEEE/CAA Journal of Automatica Sinica*, vol. 12, no. 3, pp. 485–501, 2025.

[21] L. Vlacic *et al.*, "Automation 5.0: the key to systems intelligence and industry 5.0," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 8, pp. 1723–1727, 2024.