Literature Review: Multi-Modal Multi-Agent Systems

1 Introduction

Multi-Agent Systems became a powerful solution to model and solve problems in complex and dynamic environments [9]. The pursuit of more intelligent and credible autonomous systems, akin to human society, has been a long-standing endeavor for humans. Leveraging the exceptional reasoning and planning capabilities of large language models (LLMs), LLM-based agents have been proposed and have achieved remarkable success across a wide array of tasks [19]. The research progress in multimodal learning has grown rapidly over the last decade in several areas, especially in computer vision. The growing potential of multimodal data streams and deep learning algorithms has contributed to the increasing universality of deep multimodal learning [6].

This literature review examines the current state of multi-modal multi-agent systems, exploring the technologies enabling their development, applications across various domains, and challenges that must be addressed for future advancement.

2 Multi-Modal Multi-Agent Systems: An Overview

Multi-modal and multi-agent systems are advancing toward rationality by surveying the state-of-the-art works. The emerging concern about the factual accuracy and trustworthiness of LLMs highlighting an urgent need to develop better agents or agent systems with rational reasoning processes [11]. Recent advancements in multi-modal and multi-agent frameworks offer a promising direction to address this challenge, which leverage the expertise of different agents acting together towards a collective goal.

Multiagent systems differ from single-agent systems in that several agents exist which model each other's goals and actions. From an individual agent's perspective, multiagent systems differ from single-agent systems most significantly in that the environment's dynamics can be determined by other agents [20]. Multi-agent interactions introduce challenges such as exponential growth in problem complexity due to enlarged joint action spaces and extended planning horizons, partial observability stemming from decentralized information among agents, and non-stationarity arising from concurrent agent learning processes [13].

3 Technologies and Techniques

Deep learning provides hierarchical computation models that learn multilevel abstract representations of data. There are several well-known deep architectures: convolutional neural networks (CNN), recurrent neural networks (RNN), and generative adversarial networks (GAN) [8]. Traditional multimodal data fusion methods cannot properly capture the intermodality representations and the cross-modality complementary correlations of the multimodal big data, since these are shallow models that cannot learn the intrinsic representation of data [12].

CNN architectures are used mainly as encoder parts to extract and encode data from images, and RNN architectures are used as decoder part to decode and generate captions [4]. Particularly noteworthy are neural networks, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which have revolutionized the processing of varied modalities such as visual, auditory, and textual data.

To achieve robust multimodal perception in autonomous vehicles, various fusion methods—early, late, and hybrid—integrate data from different sensors to enhance situational awareness and decision-making capabilities [7]. Neural networks, such as Convolutional Neural Networks (CNNs) for camera data and 3D convolutions

for LiDAR, extract key features from each modality, while attention mechanisms dynamically prioritize the most relevant sensor data.

4 Applications

4.1 Healthcare

Multimodal AI applications in clinical practice integrate diverse data sources such as medical images, electronic health records (EHRs), sound recordings, and genomic data to enhance diagnostic accuracy, treatment planning, and patient monitoring [15]. AI systems leveraging multiple data sources and input modalities are poised to become a viable method to deliver more accurate results and deployable pipelines across a wide range of applications [18]. The increasing availability of biomedical data from large biobanks, electronic health records, medical imaging, wearable and ambient biosensors, and the lower cost of genome and microbiome sequencing have set the stage for the development of multimodal artificial intelligence solutions [1].

4.2 Robotics and Autonomous Vehicles

Multimodal AI in autonomous driving systems processes geospatial data, video feeds, LIDAR, radar, and POI data to enable safer navigation and decision-making [3]. Autonomous vehicles rely on a variety of sensors—such as cameras, LIDAR, and radar—to navigate their environments safely. Multimodal AI systems integrate data from these different sensors, allowing the vehicle to make real-time decisions and respond to its surroundings [10].

EMMA (End-to-End Multimodal Model for Autonomous Driving), powered by Gemini, employs a unified, end-to-end trained model to generate future trajectories for autonomous vehicles directly from sensor data [21]. This research demonstrates how multimodal models can be applied to autonomous driving and explores the benefits of incorporating multimodal world knowledge.

5 Challenges and Future Directions

Despite significant advancements, several challenges hinder the widespread adoption of multimodal AI, including data fragmentation, interoperability issues, computational demands, and the need for explainable AI in clinical decision-making [17]. Real-time processing of multimodal data presents additional difficulties. Applications such as real-time translation, augmented reality, and autonomous driving require instantaneous processing and response times [16].

Alignment issues present significant challenges: ensuring that models understand prompt instructions and generate accurate, unbiased outputs requires identifying relationships between multiple modalities [5]. Co-learning difficulties in transferring knowledge across modalities remain a major hurdle. Current challenges and future research directions focus on terrain adaptability, as robots exhibit compromised locomotion on certain substrates [14].

6 Conclusion

The exploration of multi-modal multi-agent systems has illuminated their significant impact on contemporary AI applications. The multimodal interplay of various senses provides superior environmental perception and learning skills. Adapted from the human perceptual system, multimodal machine learning tries to incorporate

different forms of input and determine their fundamental connections through joint modeling [2]. Addressing the challenges identified throughout this review is crucial for enhancing the effectiveness of these systems. The successful implementation of multi-modal systems holds the potential to greatly enhance AI adaptability and intelligence in real-world situations.

References

[1] Acosta, J.N., Falcone, G.J., Rajpurkar, P., and Topol, E.J. (2022) Multimodal biomedical AI. *Nature Medicine*, 28, pp. 1773–1784.

[2] ACM Computing Surveys (2024) A Survey of Multimodal Learning: Methods, Applications, and Future. *ACM Computing Surveys*.

[3] Appen (2024) Multimodal AI. Available at: https://www.appen.com/multi-modal-ai

[4] Asghar, M.Z. et al. (2021) Recent Advances and Trends in Multimodal Deep Learning: A Review. *arXiv preprint arXiv:2105.11087*.

[5] Avasant (2024) Harnessing Multimodal AI: Innovations and Applications. Available at: https://avasant.com/report/

[6] Bayoudh, K. et al. (2021) A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38, pp. 2939–2970.

[7] European Space Agency (2024) Multimodal AI: The Future of Autonomous Vehicles. *NAVISP*.

[8] Gao, J., Li, P., Chen, Z., and Zhang, J. (2020) A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5), pp. 829–864.

[9] IEEE (2020) A Systematic Literature Review in Multi-Agent Systems: Patterns and Trends. *IEEE Conference Publication*.

[10] IMD (2024) What Is Multimodal AI and How It Works. Available at: https://www.imd.org/blog/

[11] Jiang, B. et al. (2024) Multi-Modal and Multi-Agent Systems Meet Rationality: A Survey. *Proceedings of Machine Learning Research*, 235.

[12] MIT Press (2020) A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5), pp. 829–864.

[13] Multi-agent Embodied AI (2025) Advances and Future Directions. *arXiv:2505.05108v1*.

[14] Nature Communications (2025) AI-embodied multi-modal flexible electronic robots. *Nature Communications*.

[15] PMC (2024) Advancing Clinical Practice: The Potential of Multimodal Technology in Modern Medicine.

[16] RINF Tech (2024) How Multimodal AI is Transforming Our Interaction with Technology.

[17] ScienceDirect (2025) Multimodal AI (MMAI) for next-generation healthcare: data domains, algorithms, challenges, and future perspectives.

[18] Soenksen, L.R. et al. (2022) Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5, 149.

[19] Springer (2024) A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*.

[20] Stone, P. and Veloso, M. (2000) Multiagent Systems: A Survey from a Machine Learning Perspective. *Autonomous Robots*, 8(3), pp. 345–383.

[21] Waymo (2024) Introducing Waymo's Research on an End-to-End Multimodal Model for Autonomous Driving.