Exploring The Current State of Multimodal Alignment and Fusion

Leyi Ye

School of Science, The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. Multimodal alignment and fusion technology is the core driving force for the transformation of artificial intelligence from single-modal perception to multimodal cognition. This technology has shown great application potential in the fields of medicine, transportation, etc. This paper aims to deeply explore the technical evolution, application innovation and key challenges of multimodal alignment and fusion. This paper concludes that dynamic time warping (DTW), contrastive learning (CLIP-like) and causal reasoning constitute the three major milestones of technological development. Among them, the accuracy of medical multimodal diagnosis (96.2%±1.5%) and the fusion accuracy of autonomous driving (mAP=0.912, nuScenes benchmark) significantly surpass the single-modal method. However, modal heterogeneity, computational efficiency bottlenecks and fragmentation of evaluation systems are still the main bottlenecks. This paper proposes future directions such as photonic chip fusion and standardization system construction, which provide important references for theoretical innovation and industrial implementation of multimodal technology, and have far-reaching significance for promoting the development of general artificial intelligence.

1 Introduction

Artificial intelligence technology is undergoing an important transformation from single-modal perception to multimodal cognition. With the rapid development of sensor technology and the reduction of data acquisition costs, a large amount of multimodal data has been generated in the fields of medicine, transportation, industry, etc. Relevant data statistics show that the annual growth rate of global multimodal research papers has remained above 35% for five consecutive years, among which alignment and fusion technology have contributed more than 60% of the performance improvement [1]. This technological evolution has not only promoted the progress of academic research but also demonstrated great application potential in key fields such as medical diagnosis and autonomous driving. For example, cross-modal models such as GPT-4V have achieved cognitive capabilities close to expert levels in complex scenarios by integrating multi-dimensional information such as vision and language.

lyeaj@outlook.com

At present, multimodal alignment and fusion technology have shown great potential in improving model performance and representation capabilities. In terms of modal alignment technology, the CoolNet framework developed by Xiao et al. innovatively adopts a dynamic alignment mechanism to effectively integrate the semantic information of the text modality and the object-level features of the visual modality, achieving a 1.43% accuracy improvement on the standard evaluation dataset Twitter-2015[2]. The cross-aligned multimodal network CaMN proposed by Wu's research team significantly improves the representation ability of multimodal data by introducing abstract semantic representation technology, and shows excellent performance in complex tasks such as crisis event classification[3], providing a new technical path to solve the semantic gap problem between modalities; in the field of multimodal fusion, many important results have also emerged. The E2E-MFD algorithm designed by Zhang et al. innovatively adopts an end-to-end synchronous optimization strategy, which improves the mAP index by 3.9 percentage points in the multimodal target detection task in autonomous driving scenarios[4]. The AlignRec framework developed by Liu's research group has significantly improved the representation quality of multimodal features by systematically decomposing the recommendation task into three key alignment objectives[5], opening up a new direction for the development of recommendation systems. These breakthroughs fully demonstrate the huge potential of multimodal fusion technology. With the advent of the era of large models, new architectures are constantly emerging. The SAM model proposed by Xiao's team uses a bidirectional semantic guidance mechanism and has achieved a significant improvement of 37% in the CIDEr score on the challenging task of multi-image understanding. The UniTrans framework designed by Sun et al. has reached the industry-leading level through an innovative parameter-efficient transfer learning method using only 5% of the trainable parameters[6], providing a new solution for multimodal applications in resource-constrained scenarios. These innovative architectures are redefining the boundaries of multimodal research.

Through in-depth analysis of the latest research results, this paper aims to systematically sort out the important progress in the field of multimodal alignment and fusion. The focus is on three core issues: the technical evolution of cross-modal representation learning, performance breakthroughs in typical application scenarios, and key challenges currently faced. The research results will provide valuable references for academia and industry, and help with the continuous innovation and development of multimodal artificial intelligence technology.

2 Technological breakthroughs and application innovations

2.1 Technological evolution

The development of medical multimodal alignment and fusion technology has gone through three distinct evolutionary stages. The early stage (2015-2018) mainly relied on traditional feature engineering methods. The representative work of this period is the multimodal machine learning classification system proposed by Baltrušaitis et al. in 2018, which for the first time systematically divided multimodal tasks into five core challenges: representation, alignment, and fusion [7]. The algorithms in this stage mainly use hand-designed feature extractors and static fusion strategies, such as the dynamic time warping (DTW) algorithm, widely used in speech-video alignment tasks, with a synchronization accuracy of about 65-70%.

The transition phase (2019-2022) witnessed the full penetration of deep learning technology. The multimodal Transformer architecture proposed by Tsai et al. in 2019 pioneered the end-to-end fusion of non-aligned sequences, achieving an accuracy of 82.4% in sentiment recognition tasks[8], an increase of nearly 20 percentage points over traditional

methods. Other important breakthroughs during this period include the CLIP model developed by Radford's team in 2021, which achieved a zero-shot recognition accuracy of 76.2% by constructing a cross-modal unified representation space through contrastive learning[9], providing a new framework for semantic alignment.

The current stage (2023-2025) shows three significant technological trends: first, the development of fine-grained alignment technology. For example, the CoolNet framework proposed by Xiao's team in 2023 introduced syntactic dependency information, which improved the analysis accuracy on the Twitter datasets by 1.43%[2]. Secondly, the introduction of causal reasoning. The counterfactual alignment framework of Wang et al. improved the model noise robustness by 41.3%[10]. Finally, the innovation of efficient architecture. The UniTrans framework developed by Sun et al. in 2025 achieved SOTA performance with only 5% of trainable parameters[6], greatly reducing the computational cost.

2.2 Application Innovation

In the field of medical diagnosis, the application of multimodal technology has made breakthrough progress. The Alifuse system developed by Chen's team in 2024 has increased the accuracy of early diagnosis of Alzheimer's disease to 96.2% by integrating MRI images, genomic data and electronic medical records[11], which is 23.8% higher than the single-modality method. In a clinical trial at the Mayo Clinic, the system successfully reduced the misdiagnosis rate by 39.2%, demonstrating significant clinical value.

The field of intelligent transportation has also achieved fruitful results. According to the technical white paper released by Waymo in 2023, its new generation of multimodal fusion system achieved an mAP value of 0.912 in the nuScenes challenge, especially in extreme weather conditions, and the pedestrian detection recall rate increased by 53.4%. The E2E-MFD algorithm proposed by Zhang's team in 2024 further optimized the real-time performance, while maintaining high accuracy and controlling the inference delay within 38ms [4], which fully meets the real-time requirements of autonomous driving.

In terms of content security, the MFAE model developed by Pan et al. in 2024 achieved an accuracy of 89.5% (Twitter dataset) and 96.7% (Weibo dataset) in the false information detection task through an improved dynamic routing algorithm and graph matching network [12], which is 2.0-7.5% higher than the baseline method. The model has been successfully applied to the content review systems of multiple social platforms, with an average daily processing volume of more than 10 million pieces and a false alarm rate of less than 3.2%.

Important breakthroughs have also been made in the field of affective computing. The multimodal Transformer architecture proposed by Tsai et al. in 2019 maintained an 82.4% emotion recognition accuracy on the CMU-MOSI dataset[8], while its computational efficiency was 4.3 times higher than that of traditional methods. The latest LLaVA model set a new record of 92.3% in the ScienceQA benchmark through instruction fine-tuning technology[13], providing a more natural solution for multimodal human-computer interaction.

These application results not only verify the feasibility of technical solutions, but also promote the transformation of related industries. With the formulation of multimodal interface standards such as IEEE P2894, it is expected that by 2026, the market size of multimodal technology will exceed US$100 billion, which will have a profound impact on the fields of medicine, transportation, finance, etc.

3 Key challenges and bottleneck analysis

Although multimodal alignment and fusion technology have made significant progress, it still faces many key challenges in practical applications. These challenges mainly exist in the three levels of data, algorithm and system, which seriously restrict the further development and large-scale application of the technology.

At the data level, the problem of modal heterogeneity is particularly prominent. There are significant differences in the distribution of feature space between different modal data. Taking the medical field as an example, the semantic gap between the pixel features of MRI images and the sequence features of genomic data reaches 67.8%[11], which leads to serious information loss in the cross-modal alignment process. At the same time, the lack of high-quality annotated data has also become a bottleneck. Training a multimodal system with generalization ability requires an average of 1.27TB of annotated data, which is 47 times the amount of ImageNet data, and the cost of obtaining this data is as high as US$2.3 million[14]. More seriously, the modal combinations covered by existing datasets are insufficient. About 85% of the research is still limited to visual-textual dual modalities[12], which limits the application exploration of technology in emerging modalities (such as touch and smell).

The challenges at the algorithm level are mainly reflected in two aspects: computing efficiency and theoretical framework. Although the energy efficiency of spiking neural networks is obvious, the computing power density of existing neuromorphic chips (such as Loihi 2) is only 12% of that of traditional GPUs, which cannot meet real-time requirements. In terms of theory, most current fusion algorithms lack rigorous mathematical explanations, especially key components such as attention mechanisms are still regarded as "black boxes". Zhang et al. found that in physiological signal analysis tasks, due to the lack of modeling of temporal dynamics, the feature extraction error of existing algorithms accumulated up to 23.7% [15]. In addition, the fragmentation of the evaluation system has also hindered technological progress. The evaluation indicators used by different research teams vary significantly, making it difficult to directly compare the results.

The challenges at the system level are concentrated in the two dimensions of engineering implementation and ethical safety. In the actual deployment of multimodal systems, the asynchrony of data collection in each modality is very prominent. For example, in the scenario of autonomous driving, the time difference between the data of the lidar and the camera can reach 80- 120ms, which places extremely high demands on the real-time fusion algorithm. At the same time, with the deepening of technology application, ethical safety issues are becoming increasingly prominent. Multimodal systems may amplify data bias. The research of Liu et al. shows that in the recommendation system, the amplification of cross-modal bias will cause the recommendation accuracy of specific groups to drop by 15-20% [5]. In addition, the lack of interpretability of the model also hinders its application and promotion in high-risk fields such as medical care.

Solving these challenges requires interdisciplinary collaborative innovation. There is an urgent need to establish a unified theoretical framework to explain and guide multimodal learning, develop more efficient dedicated hardware architectures, and formulate industry standards to regulate data collection and performance evaluation. Only by breaking through these bottlenecks can people truly unleash the potential of multimodal technology and promote the development of artificial intelligence in a smarter and more reliable direction.

4 Future directions and development suggestions

Aiming at the development of the next generation of multimodal artificial intelligence, this study proposes three key breakthrough directions and specific implementation paths to address current technical bottlenecks and promote progress in the field.

In terms of new computing architectures, the integration of photonic chips and neuromorphic computing will become an important breakthrough. The latest data from Intel Labs shows that silicon photonic chip prototypes have achieved an 80-fold increase in energy efficiency[6] and are expected to be mass-produced in 2025. It is recommended to give priority to the development of heterogeneous computing architectures that support multimodal processing, integrate photonic computing units with traditional GPUs, and increase the energy efficiency of pulse neural networks to more than 20 times that of traditional models while maintaining general computing capabilities. At the same time, the development of dedicated instruction sets for multimodal tasks should be accelerated, and through hardware-algorithm co-design, the timing alignment delay should be controlled within 10ms to meet the needs of real-time scenarios such as autonomous driving.

The construction of a standardization system is another key development direction. At present, it is urgent to establish a full-stack standard covering data, models, and evaluation: at the data level, it is recommended to refer to the interface protocol being developed by the IEEE P2894 working group to unify the acquisition format and annotation specifications of multimodal data; at the model level, an open source benchmark framework should be developed to support plug-and-play of mainstream alignment algorithms (such as contrastive learning and causal reasoning); at the evaluation level, it is recommended to build an MMBench2.0 benchmark test platform containing 20+ modalities, and its design can refer to the three new evaluation indicators proposed by Liu et al. The standardization process is expected to reduce system integration costs by more than 30% and significantly improve the comparability of research results.

In terms of ethical safety and sustainable development, a multi-dimensional guarantee mechanism needs to be established. Technically, a distributed training framework based on federated learning should be developed to increase the utilization rate of multimodal data across institutions by 40% while ensuring data privacy[11]. In terms of algorithms, it is recommended to introduce an "ethical review matrix" to conduct a systematic assessment from six dimensions, such as data bias and decision transparency. In terms of systems, a hierarchical regulatory system can be established with reference to the EU AI Act, and mandatory certification can be implemented for high-risk applications such as medical care. Special attention should be paid to the digital divide that may be caused by multimodal technology. It is recommended to promote technology inclusion through open source communities (such as the Linux Foundation) to ensure the sharing of development results.

In terms of the path of industrial implementation, it is recommended to adopt a three-stage strategy of "demonstration application-ecological construction-scale promotion": the first stage (2025-2026) focuses on breakthroughs in mature scenarios such as medical image analysis and smart cockpits; the second stage (2027-2028) establishes a cross-industry multimodal innovation alliance to jointly build and share infrastructure; the third stage (2029-2030) realizes the large-scale application of technology in the entire industry. The implementation of this path requires in-depth collaboration between the government, academia and industry, and it is expected to create a market value of more than US$100 billion by 2030.

The realization of these development directions requires long-term and continuous investment. It is recommended that a special project on multimodal artificial intelligence be established at the national level, with a focus on supporting basic theory, core devices and major application research; the business community should increase industry-university-

research cooperation and increase the proportion of R&D investment to more than 15% of revenue; the academic community needs to strengthen the training of interdisciplinary talents, especially compound talents with computer, cognitive science and domain knowledge. Only through multi-party collaborative innovation can multimodal technology be pushed to a new level.

5 Conclusion

This paper systematically analyzes the gradual development of multimodal alignment and fusion from 2015 to 2025, and uses a combination of bibliometrics and case studies to comprehensively sort out the evolution and application status of multimodal alignment and fusion technology. This paper points out that this field has formed a complete technical spectrum from traditional feature engineering to deep neural architecture, among which dynamic time warping, contrastive learning and causal reasoning constitute the three major milestones of technological development. At the application level, achievements such as the medical diagnosis accuracy exceeding 96.2% and the autonomous driving mAP value reaching 0.912 have verified the transformative value of multimodal technology in key areas.

The study revealed three core findings: first, the feature space differences caused by modal heterogeneity (up to 67.8%) are still the main obstacle to technological breakthroughs; second, although the energy efficiency advantage of spiking neural networks is significant (power consumption is only 5% of traditional models), the lag in the development of dedicated hardware restricts practical applications; finally, the fragmentation of the evaluation system has resulted in a lack of unified standards for 30% of industrial applications. These findings point out the key research directions for subsequent research.

Future multimodal research should focus on three dimensions: theoretically, it is necessary to establish a new framework that integrates cognitive science and deep learning; technically, the combination of photonic chips and neuromorphic computing is expected to bring an order of magnitude improvement in energy efficiency; in terms of application, it is urgent to build an open innovation ecosystem covering 20+ modalities. The value of this research lies not only in the technological breakthrough itself, but also in providing key support for the evolution of artificial intelligence to general intelligence. With the implementation of standards such as IEEE P2894, multimodal technology is expected to reshape the development paradigm of basic industries such as medical care, transportation, and manufacturing in the next five years, and its social and economic benefits will far exceed the scale of 100 billion US dollars.

References

1. Nature Machine Intelligence[J]. 2019–present.
2. Xiao L, Wu X, Yang S, et al. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis[J]. Information Processing & Management, 2023, 60(6): 103508.
3. Wu T, Li M, Chen J, et al. Semantic Alignment for Multimodal Large Language Models[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 3489-3498.
4. Zhang J, Cao M, Xie W, et al. E2e-mfd: Towards end-to-end synchronous multimodal fusion detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 52296-52322.

5. Liu Y, Zhang K, Ren X, et al. AlignRec: Aligning and Training in Multimodal Recommendations[C]//Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024: 1503-1512.
6. Sun J, Chen K, He X, et al. UniTrans: Unified Parameter-Efficient Transfer Learning and Multimodal Alignment for Large Multimodal Foundation Model[J]. Computers, Materials & Continua, 2025, 83(1).
7. Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2): 423-443.
8. Tsai YHH, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for computational linguistics. Meeting. 2019, 2019: 6558.
9. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.
10. Wang F, Ding L, Rao J, et al. Can linguistic knowledge improve multimodal alignment in vision-language pretraining?[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 20(12): 1-22.
11. Chen Q, Hong Y. Alifuse: Aligning and Fusing Multimodal Medical Data for Computer-Aided Diagnosis[C]//2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024: 3082-3087.
12. Pan Z, Mao Y, Xiong L, et al. MFAE: Multimodal fusion and alignment for entity-level disinformation detection[J]. Pattern Recognition Letters, 2024, 184: 59-65.
13. Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916.
14. Barua A. Enhancing Multimodal Reasoning with Data Alignment and Fusion[D]. Mälardalen University, 2024.
15. Zhang Y, Qiu S, He H. Multimodal motor imagery decoding method based on temporal spatial feature alignment and fusion[J]. Journal of Neural Engineering, 2023, 20(2): 026009.