Original Research Article

Open Access

Omniscient Pte.Ltd.

Research on the Practical Application of Multimodal Architecture-Based Intelligent Computing Services

Wen-Hua Zhang*

Cloud Network Technology Center, Migu Culture Technology Co., Ltd., Nanjing, Jiangsu, 210026, China

***Correspondence to:** Wen-Hua Zhang, Cloud Network Technology Center, Migu Culture and Technology Co., Ltd., Nanjing, Jiangsu, 210026, China, E-mail: 454826798@qq.com

Abstract: This paper focuses on the practical application of multimodal architectures in the field of intelligent computing services. By elaborating on key application scenarios such as audiovisual semantic understanding and image-text generation, the study explores system optimization strategies including hybrid architecture deployment and model compression and acceleration. It also clarifies the challenges faced in real-world deployment, such as data quality and cross-modal consistency, and proposes targeted solutions such as standardized annotation and dynamic verification. The research results serve as a reference for the technical implementation and industrial development of multimodal intelligent computing services, helping to address technical and managerial issues encountered in practical applications.

Keywords: Multimodal architecture; intelligent computing services; practical application; industry deployment challenges; deployment strategies

Introduction

With the acceleration of digitalization, multimodal intelligent computing services have become a key direction for improving information processing efficiency. In real-world applications, the explosive growth of multimedia data has led to a sharp increase in demands for audiovisual semantic understanding and image-text creation. However, the construction and optimization of such systems face several technical bottlenecks. Meanwhile, challenges such as data quality and regulatory compliance hinder industry-level implementation [1]. This paper adopts a grassroots perspective to analyze the technical applications and practical difficulties of multimodal intelligent computing services, aiming to explore effective solutions that can promote the adoption of related technologies across various industries.

1. Application Scenarios of Key Technologies

1.1 Audiovisual Semantic Understanding: In-depth Analysis of Multimedia Content

In the current era of rapid digital advancement, audiovisual semantic understanding presents significant development potential. With the explosive growth of multimedia content, there is an urgent need for efficient and accurate semantic interpretation of audio and video information. By integrating technologies such as speech recognition and image analysis through a multimodal architecture, it becomes possible to deeply analyze and process audiovisual

content—for instance, automatically identifying key events in videos or interpreting semantic meanings in audio clips. This capability provides strong support for fields such as intelligent surveillance and film content analysis [2].

1.2 Image-Text Generation: Breaking Barriers in Cross-Modal Creation

In practical scenarios involving image-text generation, multimodal architectures break through the limitations of traditional single-modal content creation. By leveraging the correlation between textual semantics and image features and incorporating deep learning algorithms, the system can automatically generate high-quality images based on given textual descriptions or extract key visual information from images to generate coherent textual content. This technology is widely applied in advertising design, animation production, and news illustration, significantly improving both creative efficiency and output quality [3].

1.3 Multimodal Retrieval and Analysis: Achieving Accurate Cross-Modal Search

Multimodal retrieval and analysis overcome the limitations of traditional single-modal search. In environments rich with multimedia data, users often have search needs that span multiple modalities. Multimodal architectures can integrate various types of data—including text, images, and audio—into a unified semantic space to enable fine-grained cross-modal retrieval. Whether in academic literature searches, e-commerce product searches, or multimedia archive management, this approach allows for quick and precise access to the desired information.

1.4 Human-Machine Interaction and Collaboration: Building New Models for Natural Interaction

In collaborative human-machine interaction scenarios, multimodal architectures enable more natural and efficient interaction models. Traditional interactions relying solely on keyboards and mice can hardly meet the needs of complex tasks. Multimodal interaction incorporates various input forms such as speech, gestures, and facial expressions, allowing machines to better understand human intentions and achieve smoother collaboration. This interaction model significantly enhances efficiency and user experience in domains such as industrial production, intelligent customer service, and educational training.

2. System Architecture and Performance Optimization

2.1 Hybrid Architecture Deployment Strategy

For multimodal intelligent computing services to operate efficiently, a well-designed hybrid architecture deployment strategy is essential. The computational demands of processing different modalities vary significantly—for example, video processing requires powerful graphical computation, whereas text processing relies more heavily on general-purpose processor performance. A hybrid architecture integrates cloud computing, edge computing, and local computing in a coordinated manner. Tasks requiring high real-time performance or involving sensitive data are processed at the edge or locally, reducing network latency. Meanwhile, large-scale data analysis and model training are carried out on cloud platforms that offer robust computational capabilities. This approach fully leverages the strengths of various computing paradigms, enabling flexible scheduling and efficient utilization of computing resources.

2.2 Model Compression and Acceleration Techniques

As multimodal models continue to grow in size, model compression and acceleration techniques have become critical to enhancing system performance. Multimodal architectures often contain a large number of redundant parameters and computational operations. Pruning techniques remove connections and parameters that have minimal impact on model performance, significantly reducing model size without major loss in accuracy. Quantization reduces computational complexity and memory usage by replacing high-precision floating-point operations with low-precision integer operations. Knowledge distillation transfers the knowledge learned by large, complex models to smaller, lightweight models, allowing them to maintain high accuracy while enabling faster inference—ultimately improving the responsiveness of the system.

2.3 Real-Time Inference Optimization Strategies

To achieve real-time inference in multimodal intelligent computing services, optimization must be implemented at multiple levels. During the input data preprocessing phase, techniques such as parallel processing and asynchronous loading can minimize time spent on data reading and transformation [4]. During the inference stage, the parallel computing capabilities of models

can be exploited by distributing tasks across multi-core processors or multiple GPUs for concurrent processing. Adaptive inference scheduling strategies should be developed based on the characteristics of different modalities—prioritizing critical modal information and dynamically reallocating inference resources based on system load. These measures ensure that the system continues to produce fast and accurate inference results, even in complex, dynamic application scenarios.

2.4 Energy Consumption Balancing and Control

Energy consumption is a key consideration in the operation of multimodal intelligent computing systems. The importance of energy consumption balancing and control cannot be overstated. Different computing devices and processing tasks vary greatly in their energy usage. By implementing real-time monitoring of energy consumption across system components and combining this data with task priorities and resource utilization levels, operational parameters and working modes can be adjusted dynamically. For instance, during periods of low load, the processor frequency and GPU memory bandwidth can be reduced to cut down on unnecessary energy use [5]. For tasks that are highly energy-intensive, allocating them to devices with high energy efficiency helps ensure that system performance standards are met while achieving balanced energy consumption. This not only reduces operating costs but also enhances the system's sustainability.

3. Challenges in Industry Implementation

3.1 Data Quality Management Difficulties

Multimodal intelligent computing services rely on massive datasets for model training. However, inconsistent data quality significantly hinders real-world application. Multimedia data comes from diverse sources and varies greatly in format—audio may contain background noise, videos can include blurry frames, and text data often suffers from spelling errors or semantic ambiguities. While manual annotation can enhance data accuracy, annotating multimodal data requires understanding multiple types of information simultaneously, making the process highly complex and costly. This leads to reduced annotation efficiency. Automated labeling tools struggle to handle intricate semantic relationships and often produce labeling errors. Consequently, poor data quality directly impacts model training outcomes and the performance of the final service [6].

3.2 Cross-Modal Consistency Challenges

Achieving unified semantic representation across multimodal data presents inherent challenges. Images convey meaning through color and texture, texts express abstract ideas, audio captures frequency-based characteristics, and video adds dynamic temporal content—all of which follow different modes of expression. When attempting to fuse such heterogeneous data, ensuring accurate semantic alignment across modalities becomes problematic. For instance, in describing a "beach sunset" scene, textual descriptions and visual imagery may differ in emphasis or tone. Forcing correlations between them may lead to semantic confusion. Any deviation in cross-modal consistency can cause discrepancies in model outputs, thereby undermining the reliability of the application.

3.3 Scenario Adaptability Optimization Dilemma

Demand for multimodal intelligent computing services varies widely across industry sectors. In industrial quality inspection, models must detect defects in product images with extremely high precision. In contrast, the education sector prioritizes smooth and natural human-computer interaction. Generic multimodal models often fail to meet the specialized needs of these verticals. Tailoring models to specific scenarios requires targeted optimization, which in turn depends on large training datasets and sufficient computing resources—both of which are often limited. Additionally, real-world deployment environments are complex and dynamic; for example, lighting conditions in smart security systems or background noise in voice recognition scenarios can impact model effectiveness, significantly raising the difficulty of scenario-specific optimization.

3.4 Security and Compliance Gaps

Multimodal intelligent computing services face dual challenges in data security and privacy protection. Multimedia data such as images and audio may contain biometric information, and its leakage could result in severe consequences. Cross-border and cross-platform data transmission and processing increase the risk of data theft or tampering. From a regulatory standpoint, data usage and storage laws vary by country and region, requiring enterprises to invest substantial resources

to achieve compliance. However, current security technologies and management frameworks are often insufficient to address the complex characteristics of multimodal data. Inadequate security and compliance safeguards have become a critical barrier to industry-wide implementation.

4. Industry Implementation Strategies

4.1 Standardized Data Cleaning and Annotation Procedures

To enhance the quality of multimodal data, it is essential to establish standardized cleaning and annotation workflows. In response to the issues of disorganized multimedia formats and flawed content, unified data cleaning rules can be defined—for example, applying noise reduction algorithms to eliminate background noise in audio, using image enhancement techniques to sharpen blurred images, and leveraging natural language processing tools to correct spelling and semantic errors in text. During the annotation phase, standardized templates and terminology libraries should be created to guide the annotation of multimodal data. A semi-supervised learning approach can be adopted to combine manual labeling and automated tools: human annotators handle key samples, while automation deals with large-scale initial labeling. This dual approach ensures both high-quality annotations and efficiency, providing reliable data for model training.

4.2 Cross-Modal Consistency Verification Mechanism

To address semantic mapping challenges between modalities, a dynamic verification mechanism must be implemented to ensure data consistency. First, a semantic mapping reference library for multimodal data should be developed to record standard semantic correspondences across modalities. During data fusion, contrastive learning techniques can be used to compare the multimodal data against the reference library, enabling real-time detection of semantic deviations. When significant discrepancies between image and text semantics are found, a manual review process should be triggered for expert validation and adjustment. Additionally, a feedback optimization mechanism should be incorporated to update the reference library continuously based on errors encountered in real-world applications, thereby improving the accuracy and reliability of cross-modal fusion.

4.3 Scenario-Based Model Fine-Tuning Strategy

To achieve effective scenario adaptation of multimodal models, targeted optimization should be carried out based on specific industry needs. For instance, in industrial quality inspection, where high image recognition precision is required, large amounts of product image data should be collected, and transfer learning methods used to fine-tune general multimodal models, enhancing their ability to identify defects. In the education sector, analysis of teacher-student interaction data can guide the optimization of modules such as natural language processing and gesture/emotion recognition to better suit real teaching environments. Additionally, a scenario-based evaluation system should be established, assessing models across dimensions such as accuracy and response time. Based on evaluation results, model parameters and structures can be dynamically adjusted to improve application performance in specialized contexts.

4.4 Full-Link Security and Compliance Management System

Ensuring the security and compliance of multimodal intelligent computing services requires a comprehensive full-link management system. During data storage, sensitive multimedia data—including biometric information—should be encrypted. Distributed storage technologies can be employed to decentralize data risk. For data transmission, dedicated encrypted channels should be established, and blockchain technology used to log data flow for traceability. Given regulatory differences across regions, a professional compliance team should be established to study local data usage policies and embed compliance requirements into system design and operational workflows. Periodic audits should be conducted to detect and address potential risks promptly, ensuring stable and secure operation of multimodal intelligent computing services within a compliant framework.

Conclusion

Multimodal architectures demonstrate strong potential in the field of intelligent computing services, but successful deployment requires overcoming challenges in data quality, modeling, and security. By building standardized data governance processes, implementing dynamic verification mechanisms, and establishing comprehensive end-to-end security frameworks,

coupled with scenario-specific model optimization strategies, system performance and adaptability can be significantly enhanced. With continuous technological advancement and cross-industry collaboration, multimodal intelligent computing services are expected to find deeper application across various fields, injecting new vitality into industrial transformation.

References

[1] Li Hailong, Chen Cuirong. Digital-Intelligent Transformation of Higher Education Evaluation: Value Positioning, Driving Factors, and Practical Pathways [J]. Modern Distance Education Research, 2025, 37(03): 30–39.

[2] Qin Lan, Cheng Hu, Wang Zhan. Research on Liquid Cooling Systems for Intelligent Computing Service Devices [J]. Cryogenics and Superconductivity, 2025, 53(05): 92–98.

[3] Liu Weichen, Yang Huafeng, Jiang Minmin, et al. Research on the Construction of Large Model Application Service Platforms [J]. Information Communication Technology and Policy, 2024, 50(12): 21–30.

[4] Dong Bo, Cai Shangwei. New Productive Forces in Cultural Industries: Intelligent Content Creation in Chengdu [J]. New Media Research, 2024, 10(24): 1–6+42.

[5] Long Wenxi, Ma Huiping, Wu Jun, et al. AI Implementation Scenarios, Challenges, and Countermeasures in the Power Industry [J]. Straits Science & Technology and Industry, 2024, 37(12): 21–24+28.

[6] Sun Xueyuan, Chen Yuanmou, Li Chen. Thoughts on the Development of Intelligent Computing Service Platforms in the Era of Large Models [J]. Communication World, 2023, (23): 39–40.