SILENCE

journals.panorama-sg.com/index.php/Silence

ISSN: 3054-4386
Publisher: Panorama Scholarly Group

Vol. 1, No.1, 13-39, 2026
DOI:10.5281/zenodo.18681507

UNWANTED REVIEWS

# Efficient Deployment of Multimodal Large Models: A Survey on Technical Innovations, Industrial Applications, and Challenges of Heterogeneous MoE Architecture, Low-bit Quantization, and Cloud-Edge-End Collaboration (2024-2026)

**Yaolin Zhang**[1*] **and Pengrong Huang**[2]

[1]*Jinan University, Guangzhou, China,*
[2]*South China Agricultural University, Guangzhou, China,*
[*]*Corresponding author*

Corresponding author:
18588542003@163.com

**Received:** February 12, 2026
**Accepted:** February 14, 2026
**Available online:** February 15, 2026

**Abstract**

This survey focuses on the efficient deployment of multimodal large models (MLLMs) and systematically reviews the core technologies from 2024 to 2026. At the architectural level, heterogeneous Mixture-of-Experts (MoE) architectures represented by ERNIE 4.5 achieve a balance between cross-modal parameter sharing and modality-specific computation through modality-isolated routing and router orthogonal loss, significantly improving computational efficiency. In terms of model compression and inference optimization, methods such as Layerwise Ultra-Low Bit Quantization (LUQ) and MoEQuant effectively address the quantization challenges of multimodal data, while TensorRT-LLM and ONNX Runtime provide deeply optimized inference engine solutions for hardware architectures like Blackwell. Regarding deployment strategies, cloud-edge-end collaborative architectures represented by DistMLLM, AIVD, and NANOMIND meet the requirements of low latency and high privacy while achieving cost savings and energy efficiency improvements through adaptive offloading, dynamic resource scheduling, and hardware-software co-design. Application cases show that these technologies have effectively solved industry pain points such as high latency and data privacy in medical edge diagnostic devices and quality inspection & predictive maintenance scenarios in smart manufacturing, balancing efficiency and performance.

**Keywords:** Multimodal Large Models; Heterogeneous MoE; Low-bit Quantization; Cloud-Edge-End Collaboration; Industrial Deployment

## 1. Introduction to the Technical Evolution and Current Status of Multimodal Large Model Deployment

Since 2024, multimodal large models (MLLMs) have experienced rapid evolution from theoretical exploration to large-scale industrial deployment. This chapter aims to systematically review their development context, focusing on key breakthroughs in model architecture, core deployment technologies, and application ecosystems between 2024 and 2026, laying the foundation for understanding the challenges and opportunities of current deployment.

**Cite as:** Yaolin Zhang and Pengrong Huang(2026). Efficient Deployment of Multimodal Large Models: A Survey on Technical Innovations, Industrial Applications, and Challenges of Heterogeneous MoE Architecture, Low-bit Quantization, and Cloud-Edge-End Collaboration (2024-2026). Panorama Scholarly Group | SILENCE, 1(1) (2026) 13-39. https://doi.org/10.5281/zenodo.18681507

This article is published in Silenceunder the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium,provided the original author and source are credited.

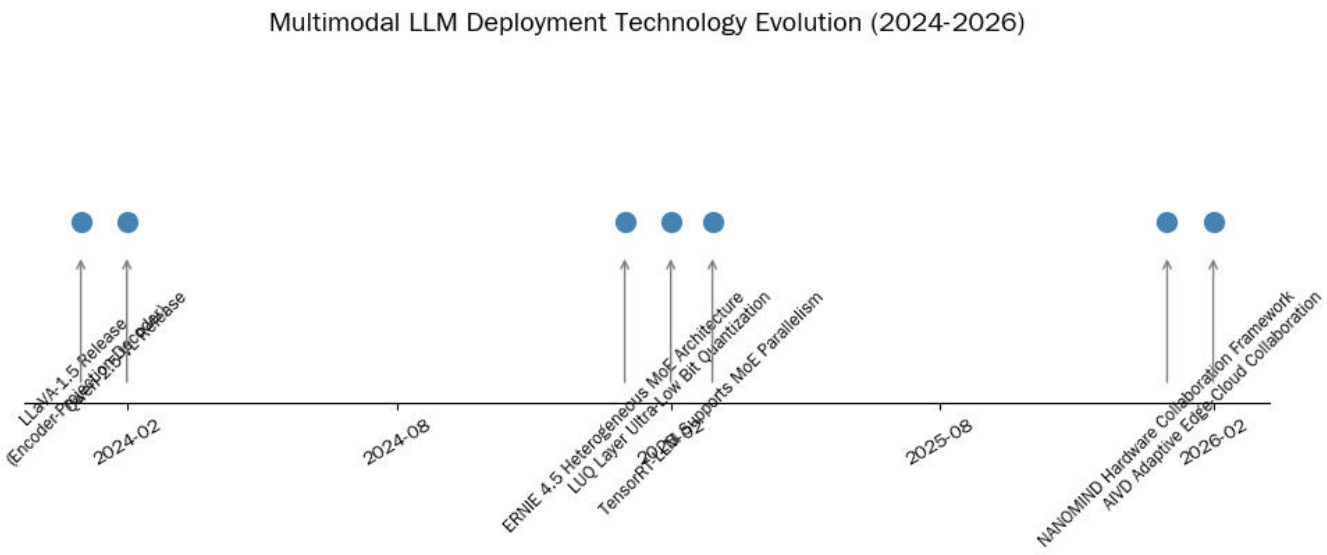



**Figure 1:** Timeline of key technological breakthroughs in multimodal large model deployment (2024–2026).

### 1.1. Architecture Evolution: From Simple Fusion to Heterogeneous Expert Mixing

The architectural design of multimodal models is the cornerstone of efficient deployment. Early vision-language models (in early 2024), such as LLaVA-1.5 and Qwen-2.5-VL, generally adopted the ”encoder-projection-decoder” paradigm [1]. Among them, visual encoders (e.g., CLIP-ViT) convert images into feature vectors, align them to the embedding space of language models through linear projection layers, and finally complete understanding and generation by large language models. Although this architecture is intuitive, it faces challenges such as low computational efficiency and insufficient multimodal fusion when processing complex multimodal interactions and long-context tasks.

In 2025, heterogeneous MoE architecture became the mainstream trend, aiming to balance performance and efficiency through dynamic and sparse computing patterns. Its core idea is to equip different modalities or tasks with specialized ”expert” sub-networks, and dynamically select the most relevant few experts for each input through a lightweight gating network. For example, Baidu’s open-source ERNIE 4.5 series models introduce a multimodal heterogeneous MoE design, achieving more refined resource allocation by sharing partial parameters between text and visual modalities while retaining dedicated experts for specific modalities [2]. This design usually includes a modality-isolated routing mechanism to ensure the specialized development of experts for different modalities, and optimizes routing decisions through technologies such as router orthogonal loss [3].

The evolution of architecture from 2025 to 2026 shows three clear directions: first, the improvement of routing reliability, which makes expert allocation for similar inputs more consistent by improving load balancing algorithms, thereby accelerating training and reducing redundancy [4]; second, the enhancement of expert diversity, with research focus shifting from simply increasing the number of experts to improving the functional differences and complementarity between experts; third, application specialization, with the emergence of MoE variants optimized for specific fields (such as medical care and industrial quality inspection), seeking the best trade-off between quality and latency [5].

### 1.2. Breakthroughs in Deployment Technologies: Quantization, Inference Optimization, and Collaborative Scheduling

To apply large multimodal models to resource-constrained edge or real-time scenarios, a series of deployment technologies have achieved revolutionary progress between 2024 and 2026.

Low-bit quantization is the key to reducing model memory footprint and computational overhead. However, the activation value distribution of multimodal models (especially visual features) often has higher variance and entropy than pure text models, making traditional uniform quantization prone to model collapse under ultra-low bits (e.g., 2 bits) [6]. The layerwise ultra-low bit quantization method proposed in 2025 specifically addresses this problem. Based on the analysis of the entropy of the activation distribution of each layer of the model, this method selectively performs aggressive quantization on more robust layers while maintaining high precision for sensitive layers. Practice shows that this method can reduce the memory footprint of the LLaVA-1.5-7B model by 40% compared to its 4-bit version, and Qwen-2.5-VL-7B by 31.5%, while maintaining a performance drop of less than 10% in benchmark tests such as MME [6]. The choice of quantization format (integer INT vs. floating-point FP) has also triggered in-depth research. Studies have shown that for 8-bit fine-grained (e.g., block size 32) quantization, MXINT8 is generally superior to its floating-point counterpart in both algorithm accuracy and hardware efficiency; but for 4-bit, floating-point formats (e.g., MXFP4) often have higher precision advantages unless supplemented with outlier suppression technologies such as Hadamard transform [7].

Inference engine optimization is another pillar for improving actual throughput. The TensorRT-LLM framework enhances native support for MoE structures, providing multiple strategies such as tensor parallelism, expert parallelism, and hybrid parallelism to efficiently utilize the computing resources of multi-GPU clusters [8]. ONNX Runtime also achieves high-performance inference on general-purpose GPU hardware through the TensorRT execution provider [9]. For the new generation Blackwell architecture, NVIDIA has further optimized quantization and speculative decoding technologies to release hardware potential [10].

Cloud-edge-end collaborative deployment strategies have moved from theory to large-scale practice. The core of such strategies is to intelligently offload computing tasks to the most suitable computing nodes (cloud, edge, or terminal). For example, the ModServe system decouples the

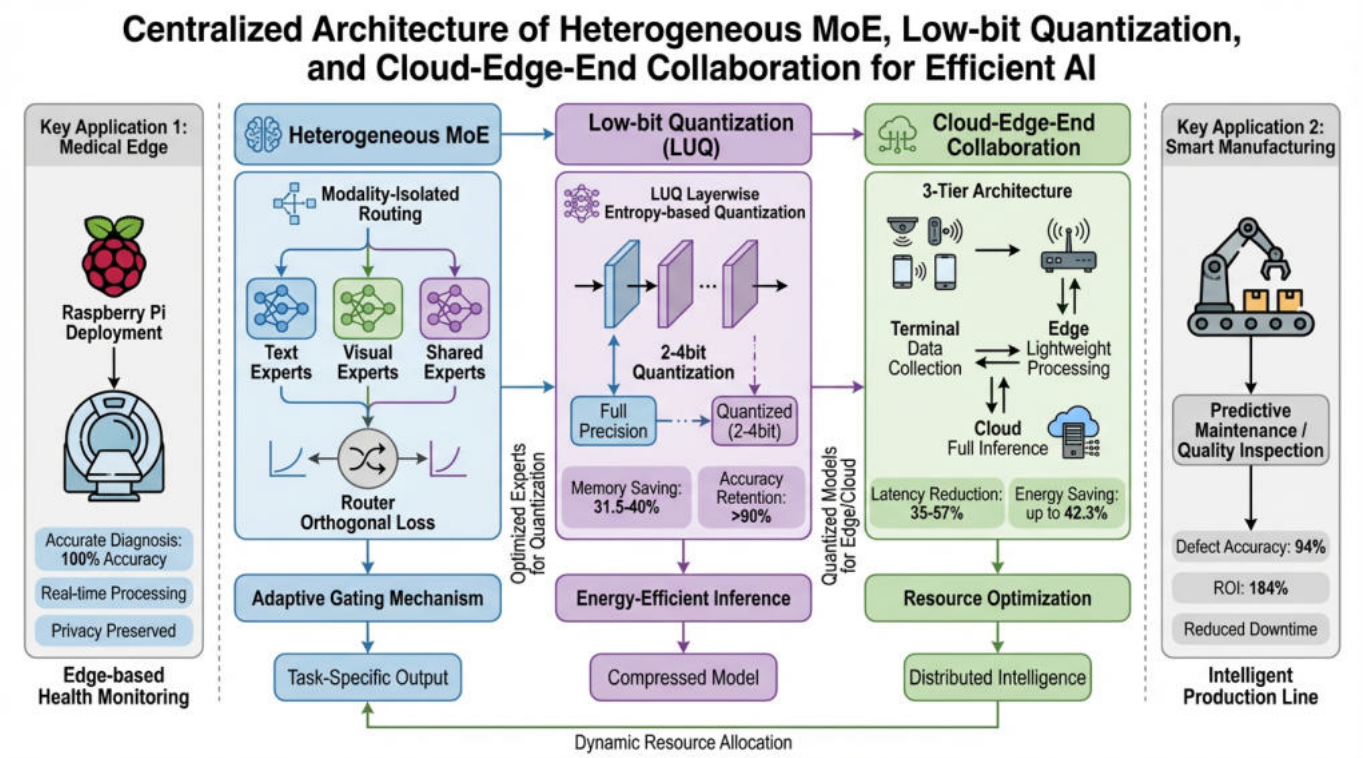



**Figure 2:** Scholarly graphical abstract for MLLM efficient deployment (2024-2026)

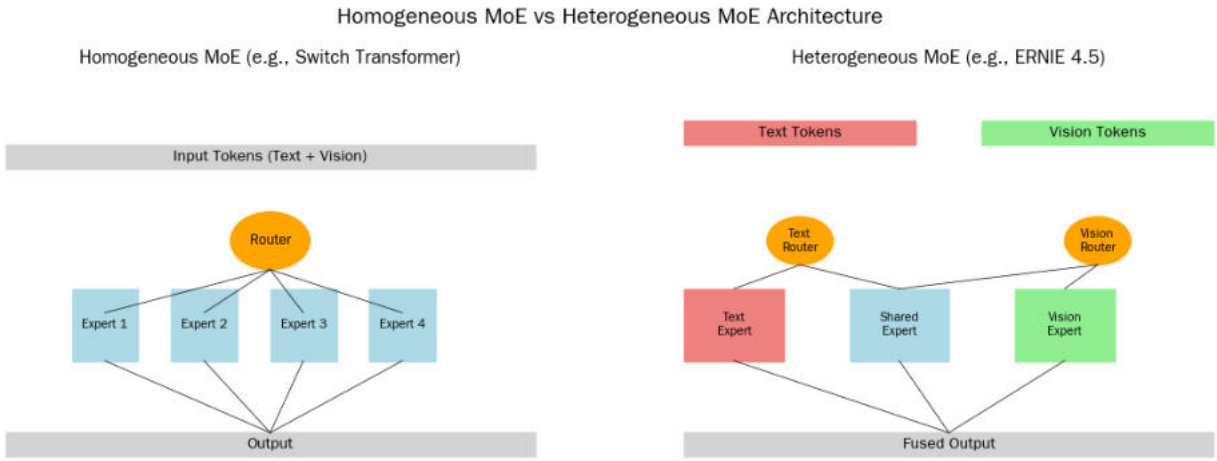



**Figure 3:** Architecture comparison between homogeneous MoE (left) and heterogeneous MoE (right).

inference pipeline of multimodal models into independent stages and realizes adaptive elastic scaling of each stage, achieving a 3.3-5.5x throughput improvement on a 128-GPU cluster while meeting the tail latency service level objective and saving 25-41.3% of costs [11]. For extremely resource-constrained scenarios, the NANOMIND framework demonstrates the power of hardware-software co-design. By splitting the model into ”modular bricks” such as vision and language and mapping them to the most matching accelerators (NPU, GPU) on the SoC, combined with dynamic load offloading and mixed quantization, it achieves a 42.3% energy reduction and up to 20.8 hours of event-triggered inference battery life on battery-powered devices [12].

### 1.3. Industry Ecosystem: Rising Adoption Rate and Specialized Applications

The industry adoption rate of multimodal large models showed exponential growth in 2025. According to an analysis of data from the Hugging Face Model Hub from June 2020 to August 2025 (covering 851,000 models and 2.2 billion downloads), the download share of multimodal generative models jumped from 1.3% in 2024 to 3.3% in 2025, with a growth rate of 3.4x [13]. The market structure has undergone significant changes: the dominant position of models from traditional U.S. tech giants has weakened, and open-source models from China (such as DeepSeek

**Table 1:** Comparison of Heterogeneous MoE and Homogeneous MoE Architectures

| Architectural Features | ERNIE 4.5 Heterogeneous MoE | Switch Transformer | DeepSeekMoE | Mixtral-8x7B |
|---|---|---|---|---|
| Expert Type | Text, Visual, Shared Experts (Heterogeneous) | Unified Expert Network (Homogeneous) | Unified Expert Network (Homogeneous) | Unified Expert Network (Homogeneous) |
| Routing Mechanism | Modality-Isolated Routing + top-k | Expert Selection Routing | Fine-grained top-k Routing | Expert Selection Routing |
| Core Loss Function | Router Orthogonal Loss + Token-Balanced Loss | Load Balancing Loss | Load Balancing Loss | Load Balancing Loss |
| Parameter Sharing Strategy | Cross-modal Sharing + Modality-Specific Parameters | No Modality Distinction | No Modality Distinction | No Modality Distinction |
| Multimodal Adaptation | Native Deep Support, Modality-Specific Experts | Limited, Need Additional Adaptation | Limited, Need Additional Adaptation | Limited, Need Additional Adaptation |
| Computational Efficiency (Example) | 47% MFU, High Inference Throughput | 40-45% MFU | 45-50% MFU | 40-42% MFU |

and Qwen) have rapidly increased their influence, occupying a significant download share in 2025 [13].
Application scenarios show a trend of deep specialization. In the field of healthcare, researchers have deployed lightweight multimodal classification models using edge computing architectures, achieving a prediction accuracy of up to 100% and an inference speed improvement of more than 28% on devices such as Raspberry Pi [14]. Specialized models based on MoE architecture (such as MedMoE) have been designed for more accurate medical vision-language understanding [5]. In the field of smart manufacturing, a predictive quality inspection framework has been successfully applied to automobile manufacturing. By integrating multiple machine learning algorithms (such as CatBoost) to perform multimodal analysis of production process and road test data, the framework increases the identification accuracy of defective vehicles to approximately 94%, and significantly increases the detected defect rate from approximately 5% in random sampling to approximately 59%, saving hundreds of thousands of minutes of production time each year [15].

### 1.4. Core Challenges and Future Directions

Despite significant progress, the deployment of multimodal large models still faces multiple challenges. The stability and efficiency of dynamic routing mechanisms are key issues; irregular memory access and cross-device communication may become performance bottlenecks [4]. The adaptation of heterogeneous computing environments requires complex scheduling algorithms to coordinate the collaborative work of hardware units with different precisions and architectures [12]. In addition, data privacy and security, model interpretability, and maintaining performance consistency in dynamic environments are all problems that must be solved for industrial implementation.
Looking to the future, technological development will deepen around the following directions: first, adaptive and context-aware deployment architectures that can dynamically adjust model configurations according to real-time resource status and task requirements; second, the deep integration of privacy-preserving computing technologies (such as federated learning) with multimodal deployment to meet compliance requirements in sensitive fields such as healthcare and finance; third, sustainability becoming a core design indicator, ranging from chip-level energy efficiency optimization to system-level carbon footprint management throughout the entire deployment life cycle. With the maturity of these technologies, multimodal large models are expected to provide efficient, reliable, and trustworthy intelligent services in a wider range of edge and terminal scenarios.

## 2. Heterogeneous Mixture-of-Experts Architecture: The Core Engine for Multimodal Adaptation and Efficient Deployment

With the deepening application of multimodal large models (MMLMs) in complex scenarios such as healthcare and manufacturing, their huge parameter size and computational requirements have become key bottlenecks restricting their widespread deployment. Traditional dense models often face challenges such as huge computing resource consumption and high inference latency when processing multi-source heterogeneous data such as text, images, and audio. Against this background, the Heterogeneous Mixture-of-Experts (Heterogeneous MoE) architecture has emerged. Through innovative technologies such as modality-isolated routing, router orthogonal loss, and multimodal token alignment, it provides a new technical paradigm for efficient and accurate multimodal inference in resource-constrained environments. The core of this architecture is to significantly reduce memory footprint and computational overhead while maintaining or even improving model performance through a sparse activation mechanism, which is the cornerstone for the large-scale and practical deployment of multimodal large models.

### 2.1. Architectural Design Principles: Paradigm Shift from Homogeneous to Heterogeneous

Traditional homogeneous MoE architectures usually adopt a unified expert network to process inputs of all modalities, and their routing mechanisms do not fully consider the essential differences in statistical distribution and semantic features between different modal data. In contrast, the heterogeneous MoE architecture has made fundamental innovations. Taking Baidu's open-source ERNIE 4.5 model family in 2025 as an example, its architectural design clearly reflects this paradigm shift [16].

This architecture explicitly divides feed-forward network (FFN) experts into three categories: text experts, visual experts, and shared experts. This division is not a simple label difference but based on a deep understanding of modal characteristics. When processing multimodal inputs, text and visual tokens are first subjected to preliminary non-routed processing by shared experts to capture cross-modal general features. Subsequently, tokens of the two modalities are independently allocated to their respective dedicated expert networks through a modality-isolated routing mechanism—text tokens are only processed by text experts, and visual tokens are only processed by visual experts [17]. This design ensures the effective extraction and deepening of modality-specific features and avoids feature interference between modalities. It is worth noting that the number of parameters of visual experts is designed to be one-third of that of text experts and is excluded in the final Transformer layer, which reflects a refined trade-off for the complexity of visual feature processing [17].

In terms of routing mechanism, ERNIE 4.5 adopts a fine-grained top-k routing strategy and introduces a router orthogonal loss of $1 \times 10^{-3}$. By forcing the weight vectors of routers to be orthogonal to each other, this loss function effectively promotes the specialized division of labor among experts, prevents ambiguity and conflicts in routing decisions, and thereby improves the overall stability and performance of the model [17]. At the same time, to address the possible increase in gradient variance during multimodal training, the architecture introduces a Token-Balanced Loss $L_{balanced}^{(i)} = \frac{1}{|U_i|+|M_i|} \sum_{j \in U_i} \ell_j^{(i)}$, which ensures the stability and efficiency of the training process through sequence length normalization of the loss [17]. The router orthogonal loss is $L_{orth} = \sum_{i=1}^{k} \sum_{j=1}^{k} (\hat{w}_i^T \hat{w}_j - \delta_{ij})^2$, with a coefficient of $1 \times 10^{-3}$.

## 2.2. Performance Advantages: Dual Improvement in Efficiency and Accuracy

The performance advantages brought by the heterogeneous MoE architecture have been fully verified in the measured data of ERNIE 4.5. In terms of training efficiency, its largest language model achieves a model FLOPs utilization (MFU) of up to 47% on a cluster of 2016 NVIDIA H800 GPUs, demonstrating excellent hardware utilization capabilities [17]. In terms of inference performance, a single H800 node achieves a throughput of 56k input tokens per second (TPS) and 18k output TPS [17].

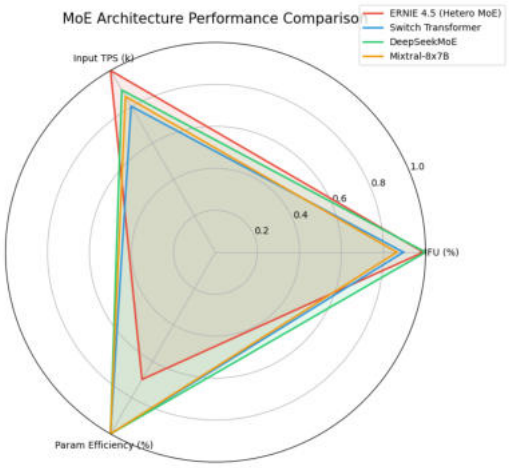



**Figure 4:** Compare MFU, throughput, and parameter efficiency across MoE architectures.

More importantly, this architecture performs outstandingly in parameter efficiency. For example, the total number of parameters of the ERNIE-4.5-21B-A3B-Base model is 21 billion, which is about 70% of the Qwen3-30B model, but it shows better performance in multiple mathematical and reasoning benchmark tests [16]. This indicates that the heterogeneous MoE design can achieve equivalent or even stronger capabilities with fewer activated parameters, which is crucial for reducing deployment costs.

To more intuitively compare the differences between heterogeneous MoE and mainstream homogeneous MoE architectures, the following table summarizes them from multiple dimensions:

As can be seen from the table, the core advantage of heterogeneous MoE lies in its design natively optimized for multimodal scenarios. Modality-isolated routing and expert type division enable it to process data of different modalities more accurately without relying on subsequent adaptation layers that may bring performance losses like homogeneous architectures.

## 2.3. Key Adaptation Technologies in Resource-Constrained Environments

The efficiency of the heterogeneous MoE architecture is particularly important in resource-constrained environments such as edge computing and mobile devices. Its adaptability is mainly achieved through the following key technologies:

1. Deep Integration with Low-bit Quantization: Due to their sparsity and dynamicity, MoE models pose unique challenges to quantization. Studies have pointed out that MoE quantization faces two major problems: inter-expert imbalance (large differences in activation frequencies between different experts) and intra-expert imbalance (large differences in the importance of different channels within the same expert) [18]. To address this, advanced quantization schemes such as MoEQuant propose an Expert-Balanced Self-Sampling (EBSS) method, which guides the construction of the calibration set by using the cumulative probability of tokens and expert balance indicators to ensure balanced expert loads after quantization; at the same time, it adopts Affinity-Guided Quantization (AGQ), which incorporates the affinity between samples and experts into the quantization process to more accurately evaluate the impact of samples on the weights of specific experts [18]. These technologies

**Table 2:** Performance Comparison of Quantization Methods on Multimodal Models

| Model | Quantization Method | Average Bit Width | Memory Saving | MME Perception Score | MME Cognition Score | TextVQA Accuracy |
|---|---|---|---|---|---|---|
| LLaVA-1.5 7B | GPTQ 4-bit | 4.00 | Baseline | 1450 | 347 | 56.8 |
| LLaVA-1.5 7B | LUQ 16 Layers | 2.54 | 40% | 1365 | 257 | 53.4 |
| Qwen 2.5 VL 7B | GPTQ 4-bit | 4.00 | Baseline | 1638 | 610 | 84.2 |
| Qwen 2.5 VL 7B | LUQ 12 Layers | 2.75 | 31.5% | 1640 | 600 | 81.9 |

enable heterogeneous MoE models to maintain high task accuracy under 4-bit or even 2-bit ultra-low bit quantization. For example, applying the Layerwise Ultra-Low Bit Quantization (LUQ) strategy to the LLaVA-1.5 model can reduce memory footprint by 40% compared to the 4-bit model, while controlling the performance drop within 10% [6].

2. Support for Modular Inference and Hardware Co-design: The heterogeneous MoE architecture is naturally compatible with the idea of modular inference frameworks. For example, the NANOMIND framework decomposes large multimodal models into independent "modular bricks" such as visual encoders and language models, and dynamically offloads each module to the most suitable heterogeneous accelerator (such as NPU, GPU) for execution [12]. Text experts and visual experts in heterogeneous MoE can be mapped to different hardware units respectively, realizing fine-grained division and parallel execution of computing tasks. This hardware-software co-design has achieved remarkable results on battery-powered edge devices, achieving a 42.3% reduction in energy consumption and nearly 20.8 hours of continuous inference capability [12].

3. Empowering Resource-Aware Optimization in Federated Learning: In federated learning scenarios that protect data privacy, the heterogeneous MoE architecture also shows great potential. The HFedMoE framework dynamically selects the most critical subset of experts for local fine-tuning from the perspective of information bottleneck according to the computing budget of each client (such as different hospitals), instead of updating the entire large model [19]. This resource-aware scheduling strategy maintains model performance (even if half of the clients are resource-constrained, it can still maintain an accuracy of over 95%) while reducing the average GPU usage rate by 10%, providing an efficient and feasible path for collaborative training of multimodal AI models across institutions [19].

### 2.4. Application Prospects and Future Directions

With its excellent efficiency and flexibility, the heterogeneous MoE architecture has become a key technology driving the industrial implementation of multimodal large models. Looking to the future, its development may focus on the following directions: deepening cross-modal expert knowledge transfer mechanisms to reduce reliance on large-scale annotated data; further optimizing adaptive routing algorithms to cope with more complex and variable input data; hardware-aware Neural Architecture Search (NAS) to customize optimal expert configurations and routing strategies for specific chips; and enhancing model interpretability to establish user trust in AI decisions through visualizing expert activation and routing decisions.

In summary, through its sophisticated modality-aware design, the heterogeneous mixture-of-experts architecture has successfully achieved the best balance between computational efficiency and model performance. It not only provides a practical solution for the current deployment of multimodal large models in edge-side and cost-sensitive scenarios but also points out an important technical evolution direction for the construction of the next generation of efficient and general artificial intelligence systems.

## 3. Low-bit Quantization and Inference Engine Optimization: Technical Principles and Industrial Practice

The application of multimodal large models in fields such as medical image analysis and industrial visual inspection faces huge challenges in computing and memory overhead. To achieve efficient deployment on resource-constrained edge devices or high-throughput cloud environments, low-bit quantization technology and inference engine optimization have become core solutions. This chapter will systematically analyze 2-bit/4-bit low-bit quantization technologies for multimodal large models, focus on interpreting selective quantization methods based on hierarchical entropy perception (such as LUQ), and conduct an in-depth analysis of the optimization strategies and deployment practices of TensorRT-LLM and ONNX Runtime on advanced hardware architectures such as NVIDIA Blackwell.

### 3.1. Challenges and Innovative Methods of Low-bit Quantization for Multimodal Large Models

Compared with pure text models, the quantization of multimodal large models (MLLMs) faces unique challenges. Studies have shown that tokens generated by non-text modalities such as images and their intermediate layer activations have significantly higher statistical variance and entropy than text tokens, resulting in lower tolerance for ultra-low bit (e.g., 2-bit) quantization [6]. However, different layers of the model have significant differences in quantization sensitivity. The core insight of the LUQ (Layerwise Ultra-Low Bit Quantization) method is that by analyzing the entropy of the activation distribution of each layer, it is possible to identify "robust" layers with more concentrated activation distributions that are more resistant to quantization accuracy loss [20].

The technical architecture of LUQ adopts a four-step process: first, generate multimodal calibration tokens through an enhanced vision-language model; second, extract intermediate layer activations of the model; third, use K-means clustering to estimate the entropy value ($H_i$) of the activation distribution of each layer, whose calculation formula is $H_i = -\sum_{k=1}^{K} P_i(k) \log P_i(k)$, where $P_i(k)$ is the probability that the token is assigned to the k-th cluster center; finally, implement iterative quantization, starting from the layer with the lowest entropy, gradually applying post-training quantization of ultra-low bits (such as the BiLLM method) to it, while maintaining higher precision (such as 4-bit GPTQ) for other

layers until the preset performance threshold or memory budget is met [20].

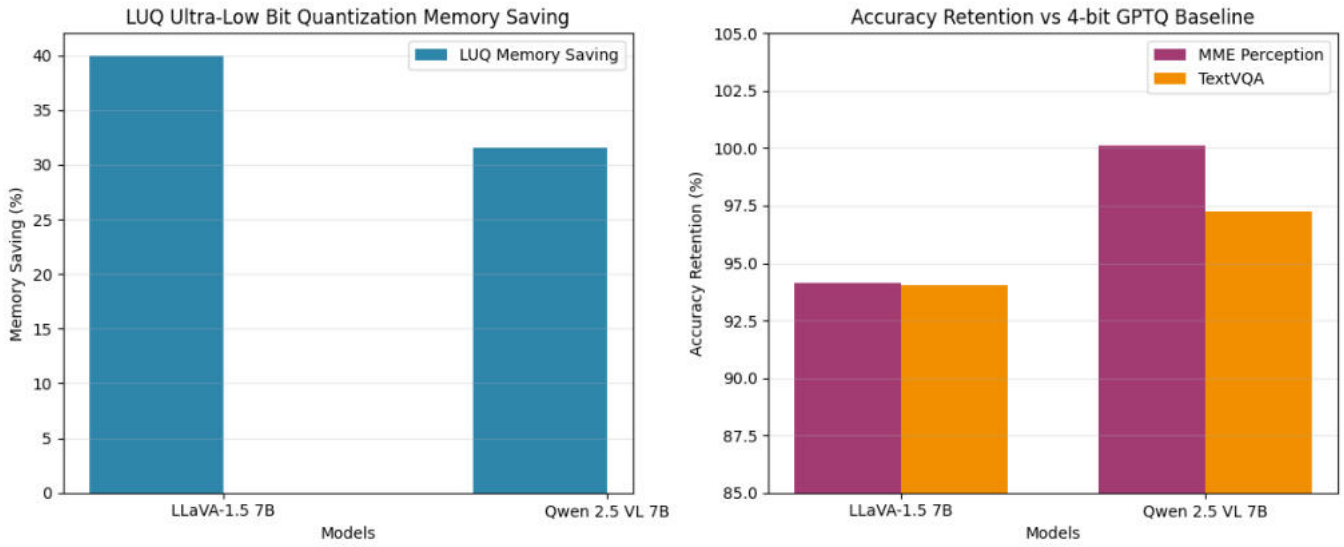



**Figure 5:** Visualize memory saving and accuracy retention of GPTQ (4-bit) vs. LUQ (ultra-low bit) on LLaVA-1.5 and Qwen 2.5 VL models.

The mixed modal calibration strategy is another key innovation of LUQ. When constructing the calibration dataset, this method does not only use text tokens but samples text tokens and multimodal (such as image) tokens in proportion, i.e., $D_{cal} = \text{Sample}(T, (1-\alpha)N) \cup \text{Sample}(M, \alpha N)$, where $\alpha$ controls the proportion of image tokens. This strategy significantly improves the ultra-low bit quantization performance of the model on Visual Question Answering (VQA) tasks [6].

In terms of performance, LUQ has achieved breakthroughs on multimodal models. For the LLaVA-1.5 7B model, LUQ achieves an average of 2.54 bits per parameter while reducing memory footprint by 40%; for the Qwen 2.5 VL 7B model, it achieves a 31.5% memory saving with an average bit width of 2.75 [20]. In terms of precision retention, the MME perception score of LLaVA-1.5 only drops by 5.8%; the precision drops of Qwen 2.5 VL on complex benchmarks such as DocVQA and POPE are only 5.5% and 1.4% respectively [6].

Quantization for MoE architecture faces additional challenges, as its sparse activation and dynamic routing characteristics will exacerbate quantization errors. The MoEQuant method effectively improves the low-bit quantization accuracy of MoE models by introducing Expert-Balanced Self-Sampling (EBSS) to construct a calibration set with balanced expert activation distribution and Affinity-Guided Quantization (AGQ) to prioritize protecting weights crucial for routing decisions [18]. On DeepSeekMoE-16B, 4-bit quantization of MoEQuant increases the HumanEval benchmark accuracy by more than 10 percentage points and the inference speed by more than 1.2x [18].

### 3.2. Deep Optimization of TensorRT-LLM on Blackwell Architecture

NVIDIA's TensorRT-LLM is a high-performance optimization library for large model inference, which has achieved extreme hardware-software co-optimization on the latest Blackwell architecture (such as B200 GPU) [21].

Native support for FP4 quantization is a key feature of the Blackwell architecture. TensorRT-LLM can seamlessly load model weights in FP4 format and automatically call computing kernels optimized for this format, while ensuring inference accuracy, significantly reducing memory bandwidth requirements and computational overhead [21].

Expert parallelism is a key technology for the efficient deployment of MoE models on multiple GPUs. TensorRT-LLM supports three parallel modes: tensor parallelism (TP), expert parallelism (EP), and a hybrid mode of the two [8]. Tensor parallelism distributes weight slices of each expert to different GPUs; expert parallelism distributes complete experts to different GPUs; the hybrid mode combines the advantages of both, performing parallelism both between and within experts to more evenly balance the load and improve overall efficiency [8]. This design is crucial for deploying large MoE models such as Mixtral 8×7B.

Dynamic offloading and kernel fusion technologies further release hardware potential. For example, dynamic management of the key-value cache of the attention mechanism through KV cache offloading optimizes memory usage [21]. More importantly, for long-sequence inference, TensorRT-LLM significantly improves throughput through optimized kernels such as Multiblock Attention. Data shows that on H200 GPU, for long sequences, its throughput can be increased by more than 3x [22]. Technologies such as speculative decoding have also been integrated to further improve the throughput of a single inference [22].

Support for multimodal models is continuously expanding. TensorRT-LLM has begun to experimentally support the optimization of image generation models (such as the Stable Diffusion series), significantly accelerating the generation process by integrating model optimizers, supporting quantization-aware training and low-rank adaptation methods [22].

### 3.3. Cross-Platform Quantization and Deployment Framework of ONNX Runtime

As a cross-platform inference engine, the quantization toolchain of ONNX Runtime provides a standardized path for deploying multimodal models on multi-vendor hardware [23].

Quantization representation formats are mainly divided into two types: operator-oriented (QOperator) and tensor-oriented (QDQ). The QOperator format provides a dedicated ONNX definition for each quantization operator (such as QLinearConv); while the QDQ format simulates the

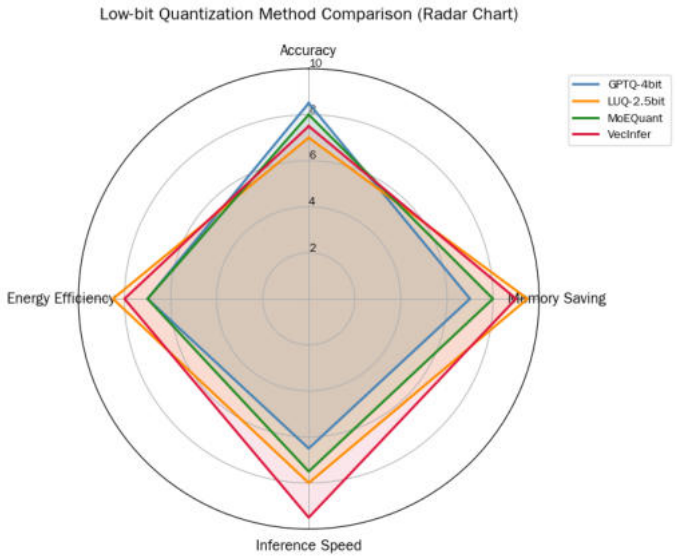



**Figure 6:** Radar chart comparing four low-bit quantization methods across four performance dimensions.

quantization and dequantization process by inserting QuantizeLinear and DeQuantizeLinear nodes between original operators, which is more flexible and framework-compatible [23].

The choice of quantization method depends on the model type. For models with Transformer architecture (including multimodal large models), dynamic quantization is usually recommended because their activation range changes greatly, and dynamically computing quantization parameters can better maintain accuracy [23]. For CNN models, static quantization that requires a calibration dataset is more suitable to obtain higher inference efficiency [23].

Int4/UInt4 weight quantization is an important progress of ONNX Runtime. It supports block-level (e.g., 128) 4-bit quantization of weights for operators such as MatMul and Gather, significantly compressing the model size. This requires the model ONNX operator set version to be no lower than 21 [23].

GPU quantization integration is mainly achieved through the TensorRT execution provider. Developers need to provide calibration data, and ONNX Runtime calculates and saves quantization parameters, and finally loads and runs the quantized model through TensorRT EP, thereby obtaining acceleration on GPUs supporting Tensor Core (such as T4, A100) [23].

### 3.4. In-depth Comparison of INT and FP Quantization Formats and Hardware Collaboration

With the tilt of AI hardware (such as Blackwell) towards low-precision floating-point formats, clarifying the advantages and disadvantages of integer (INT) and floating-point (FP) formats is crucial for algorithm-hardware co-design. A systematic study has revealed key performance crossing points [7].

Algorithm accuracy varies at different bit widths. For 8-bit quantization, MXINT8 is comprehensively superior to MXFP8 in direct conversion inference. For 4-bit quantization, the situation is more complex: without outlier mitigation technology, FP formats (such as MXFP4) usually have accuracy advantages; but when technologies such as Hadamard transform are applied, NVINT4 can surpass NVFP4 [7]. Theoretical analysis shows that this advantage depends on the crest factor of the data, and Hadamard transform can effectively reduce this factor, making INT formats superior in more scenarios [7].

Hardware efficiency is distinctly different. Under the premise of achieving the same throughput, fine-grained INT formats are significantly superior to their floating-point counterparts in area and energy consumption. For example, in a single-format comparison, MXINT8 reduces energy consumption by 37% compared to MXFP8; in a mixed-format configuration, "MXINT8+NVINT4" reduces the area by approximately 34% compared to "MXFP8+NVFP4", which benefits from the simpler circuit reuse of the INT computing pipeline [7].

These findings challenge the simple perception that "floating-point formats are always optimal", emphasizing the importance of considering INT formats in the algorithm design stage according to specific tasks, models, and hardware characteristics to achieve the best balance of accuracy, power consumption, and efficiency [7].

### 3.5. Quantization Optimization Practice in Edge Computing Scenarios

In edge scenarios with strictly limited resources and power consumption, quantization technology directly determines the feasibility of multimodal AI applications.

The NANOMIND framework is a model of hardware-software co-design. It decomposes large multimodal models into modular "bricks" such as visual encoders and language models, and dynamically schedules them to the most suitable accelerators (such as NPU, GPU) in the SoC according to the computing characteristics of each module [12]. Through customized low-bit GEMM computing kernels and token-aware buffer management, the framework achieves a 42.3% reduction in energy consumption and 11.2% saving in GPU memory on battery-powered devices, enabling the LlaVA-OneVision model to run with continuous event-triggered inference for nearly 20.8 hours [12].

Cases of medical edge diagnosis further confirm the value of quantization optimization. A study deployed a lightweight CNN optimized by genetic algorithm on multiple Raspberry Pi edge nodes, achieving a prediction accuracy of up to 100% and an inference speed improvement of

more than 28% in multimodal medical data classification tasks, significantly reducing computational complexity [14]. This demonstrates the dual advantages of edge intelligence in improving diagnosis efficiency and protecting data privacy.

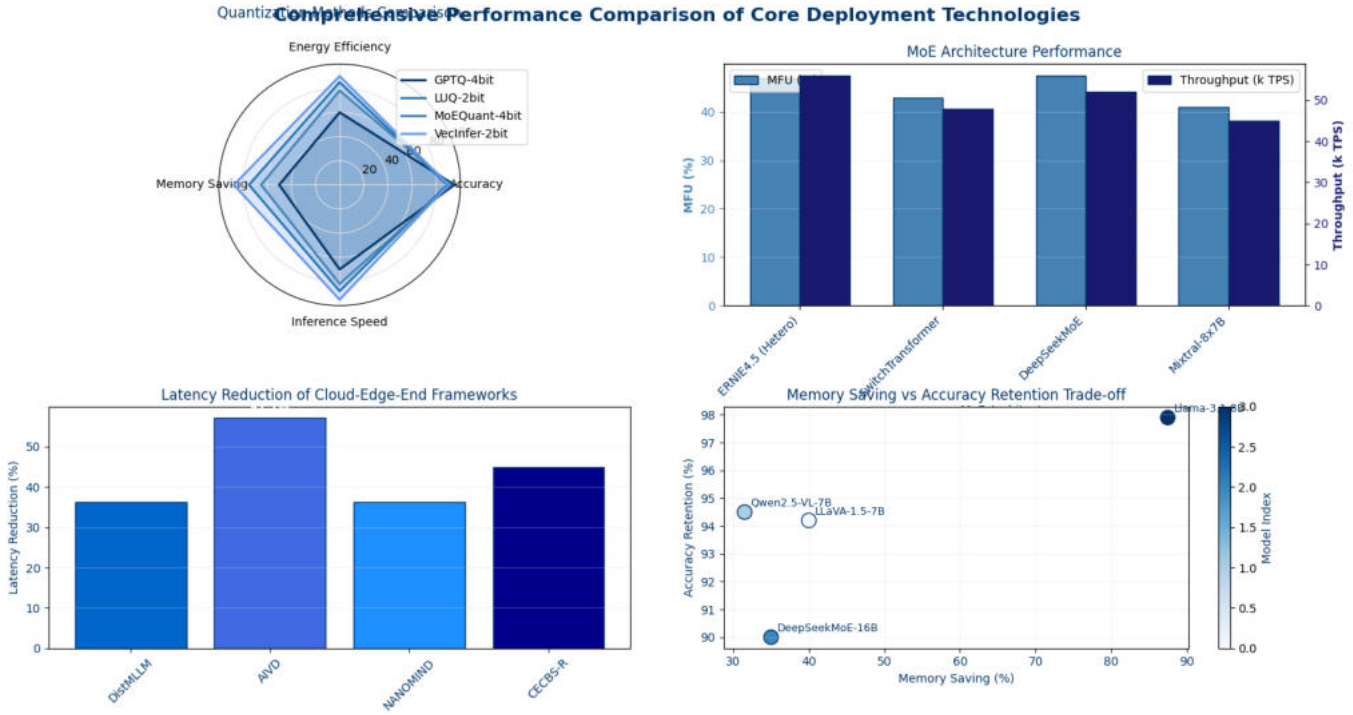



**Figure 7:** Comprehensive performance comparison of key MLLM deployment technologies. (a) Multi-dimensional radar chart of quantization methods; (b) MFU and throughput comparison of MoE architectures; (c) Latency reduction of cloud-edge-end collaborative frameworks; (d) Memory saving vs. accuracy retention trade-off of quantization schemes

KV cache compression is crucial for processing long sequences. The VecInfer method performs 2-bit quantization on the KV cache of the Llama-3.1-8B model through vector quantization technology. At a sequence length of 196k, its customized CUDA kernel achieves a 2.7x acceleration in self-attention computation and an 8.3x reduction in end-to-end decoding latency, while the average accuracy only drops by 2.1%, maintaining performance comparable to full-precision in complex mathematical reasoning tasks [24].

### 3.6. Technical Trends and Future Outlook

Based on current progress, low-bit quantization and inference engine optimization technologies are moving towards more refined, adaptive, and standardized directions:

1. Mixed Precision and Adaptive Quantization: Future quantization strategies will be more dynamic, able to adaptively adjust the bit width of different parts during inference according to the complexity of input data, the sensitivity of model layers, and real-time energy consumption constraints, achieving a better accuracy-efficiency trade-off.
2. Deep Hardware-Algorithm Collaboration: For new hardware features such as FP4 of Blackwell and specific instruction sets of NPU, more dedicated quantization algorithms and fused kernels will emerge to release hardware computing power from the bottom layer.
3. Specialization of Multimodal Quantization: Developing dedicated quantization calibration strategies and error compensation methods for the data characteristics and task requirements of different modalities such as vision, speech, and text will become the key to improving the end-to-end performance of multimodal models.
4. Standardization and Ecosystem Integration: Promoting the interoperability between different technology stacks such as the QDQ format of ONNX Runtime and the FP4/INT4 support of TensorRT, and reducing the migration cost of models from training to cross-platform deployment are prerequisites for large-scale industrial applications.

With the maturity and integration of these technologies, multimodal large models will be able to be deployed more efficiently and widely in various scenarios from cloud data centers to portable medical devices and industrial quality inspection terminals, truly becoming the core driving force for promoting industrial intelligence upgrading.

## 4. Cloud-Edge-End Collaborative Deployment Strategy and Resource Scheduling Logic

With the widespread application of multimodal large models (MLLMs) in fields with high real-time and privacy requirements such as healthcare and smart manufacturing, the traditional centralized cloud deployment model faces severe challenges in latency, bandwidth, and cost. The cloud-edge-end collaborative deployment architecture achieves the best balance of performance, efficiency, and security by intelligently distributing and scheduling computing tasks among cloud, edge nodes, and terminal devices. This section will deeply analyze the core architectural paradigms of multimodal large models in collaborative deployment, interpret advanced frameworks represented by DistMLLM, AIVD, and NANOMIND and the mathematical principles of their resource scheduling algorithms, and focus on exploring their practical effects and performance-cost trade-offs in two typical scenarios: medical care and industry.

### 4.1. Architectural Design Paradigms of Collaborative Deployment

Current mainstream cloud-edge-end collaborative architectures mainly focus on three core design ideas: task decoupling, hierarchical collaboration, and software-hardware collaboration, aiming to fully utilize the advantages of computing resources at different levels.

Task Decoupling Execution Architecture, represented by the DistMLLM framework, its core innovation lies in decomposing the complex multimodal inference process into two independent stages: encoding and inference [25]. In this architecture, edge devices with weak computing capabilities (such as Jetson AGX Orin) are responsible for receiving raw multimodal data (such as images, audio) and performing efficient encoding using lightweight GPUs to generate compact tensor representations. These low-dimensional tensors are then transmitted to edge nodes or the cloud with stronger computing capabilities (such as NVIDIA A100) for final large language model inference. This decoupling strategy significantly reduces the amount of data to be transmitted, thereby effectively reducing network latency, and optimizes overall resource utilization by offloading computationally intensive tasks to more suitable hardware.

Three-Tier Collaborative Detection Architecture, taking the AIVD framework as an example, constructs a refined pipeline collaboration system [26]. This architecture includes three logical layers: 1) Edge Perception Layer, composed of heterogeneous devices deployed with lightweight object detection models (such as YOLOv12s), responsible for high-frequency and low-cost local data collection and preliminary positioning; 2) Cloud Semantic Layer, hosting complete MLLMs (such as Qwen2-VL-7B), performing fine-grained classification and structured semantic generation on the positioned areas uploaded by the edge; 3) Dynamic Scheduling Layer, acting as a global coordinator, intelligently distributing computing tasks across layers based on real-time monitored device status, network bandwidth, and task queues. This design realizes the separation and collaboration of perception and decision-making, balancing the real-time performance of the edge and the strong cognitive ability of the cloud.

Hardware-Software Co-design Architecture, exemplified by the NANOMIND framework, its core idea is to decompose large multimodal models into modular ”bricks” such as vision, language, and audio, and dynamically offload each module to the most ideal accelerator for execution according to the characteristics of different computing units (GPU, NPU, DSP) in the Unified Memory Architecture (UMA) SoC [12]. Through customized low-bit GEMM computing kernels, system-level scheduling strategies, and driver optimization for specific hardware, the framework realizes the operation of complete MLLM inference on battery-powered devices with harsh resources, providing extreme energy efficiency ratio for edge deployment [12].

### 4.2. Mathematical Principles and Implementation of Resource Scheduling Algorithms

The core of efficient collaborative deployment lies in an intelligent scheduler that can dynamically adapt to network fluctuations and hardware heterogeneity. Current advanced scheduling algorithms are generally based on multi-objective optimization theory, seeking Pareto optimality among latency, energy consumption, accuracy, and cost.

The heterogeneous resource-aware dynamic heuristic scheduling algorithm has been successfully applied in the AIVD framework [26]. Its core is to construct a node state scoring function. The algorithm periodically collects and normalizes four key indicators for each edge node to form a state vector $R_i = [U_i, Q_i, B_i, L_i]^T$, where $U_i$ represents the CPU idle rate, $Q_i$ is the queue congestion score, $B_i$ is the available bandwidth capacity, and $L_i$ is the network latency. The comprehensive score $S_i$ of each node is calculated through a linear fusion function with adjustable weights: $S_i = w^T R_i = \alpha U_i + \beta Q_i + \delta B_i + \varepsilon L_i$, where $w = [\alpha, \beta, \delta, \varepsilon]^T$ is a hyperparameter that can be adjusted according to business objectives (such as prioritizing low latency or high throughput). To smooth score fluctuations and enhance temporal stability, the algorithm further adopts an exponential smoothing strategy to dynamically update the score: $S_i(t+1) = \eta S_i(t) + (1-\eta) \cdot f(R_i(t))$, where $\eta$ controls the fusion ratio of historical information and current information. Finally, each newly arriving task is assigned to the node with the highest score $i^* =_{i \in N} S_i$, thereby achieving adaptive load balancing [26].

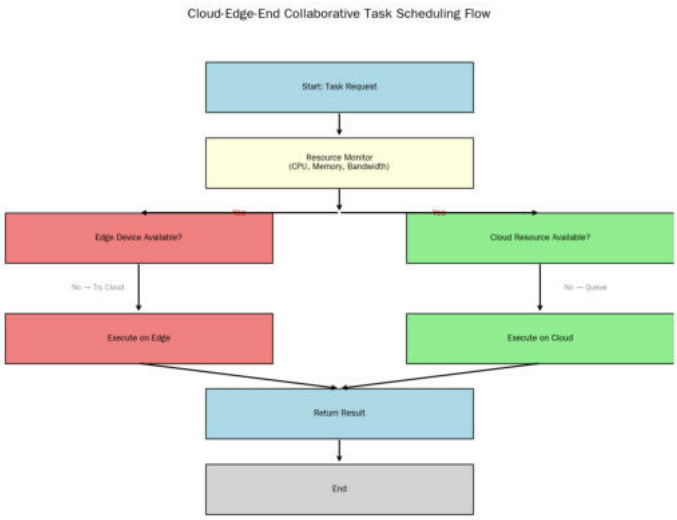



**Figure 8:** Task scheduling flow in cloud-edge-end collaborative deployment.

The multi-agent bandit algorithm is adopted by the DistMLLM framework, whose design goal is to maximize the long-term total profit of all service providers under uncertain device status and competing service provider interests [25]. This algorithm is essentially an online learning process, divided into two stages: exploration and exploitation. In the exploration stage, each service provider randomly assigns encoding and

inference subtasks to different edge workers to collect sample data on task execution time, energy consumption, and rewards. In the exploitation stage, the algorithm determines an optimal task allocation mapping by solving a multiple knapsack optimization problem based on the learned performance model to maximize expected profit. This mechanism can continuously learn and optimize decisions in dynamic environments, achieving effective resource allocation [25].

Resource recommendation and scheduling based on machine learning are reflected in the CECBS-R framework, which uses artificial neural networks (ANN) for runtime estimation and resource recommendation to optimize task scheduling costs in cloud-edge environments [27]. Its core components include a recurrent neural network (RNN) for server energy consumption control, a multi-layer perceptron (MLP) for task runtime classification, and a long short-term memory network (LSTM) for recommending the optimal execution environment (cloud or edge). Through the learning of historical data by machine learning models, the system can provide intelligent decisions for job scheduling, and experiments show that the framework can achieve about 65% cost savings [27].

### 4.3. Quantitative Verification of Performance-Cost Trade-off

The actual effects of different collaborative deployment strategies need to be quantitatively evaluated through key performance indicators. Based on experimental data from existing frameworks, the trade-off relationships between various strategies in terms of latency, accuracy, energy consumption, and cost can be clearly seen.

In terms of latency and throughput, the AIVD framework shows significant advantages in industrial visual inspection scenarios. In a 16-node experiment simulating high network pressure (500ms latency), its scheduling strategy improves throughput by 11.1% and 14.6% compared to Round-Robin (RR) and Static Resource-Aware (SRA) baselines, respectively, and reduces average latency by more than 30% [26]. The DistMLLM framework reduces end-to-end latency by 36.2% through decoupled transmission [25].

In terms of energy efficiency and battery life, the NANOMIND framework performs particularly outstandingly. Through software-hardware co-optimization, its energy consumption is reduced by 42.3% compared to traditional implementations, and GPU memory usage is reduced by 11.2% [12]. In the low-power on-demand cascaded inference mode, a device equipped with a 2000mAh battery can run multimodal models such as LLaVA-OneVision for nearly 20.8 hours, which is crucial for mobile medical devices [12].

In terms of accuracy retention, the visual-semantic collaborative enhanced fine-tuning strategy of the AIVD framework increases the average classification accuracy of the InternVL3.5 model on the PCB defect dataset to 0.972, far exceeding the zero-shot baseline [26]. The DistMLLM framework also achieves a prediction accuracy of up to 99.09% in specific tasks [25].

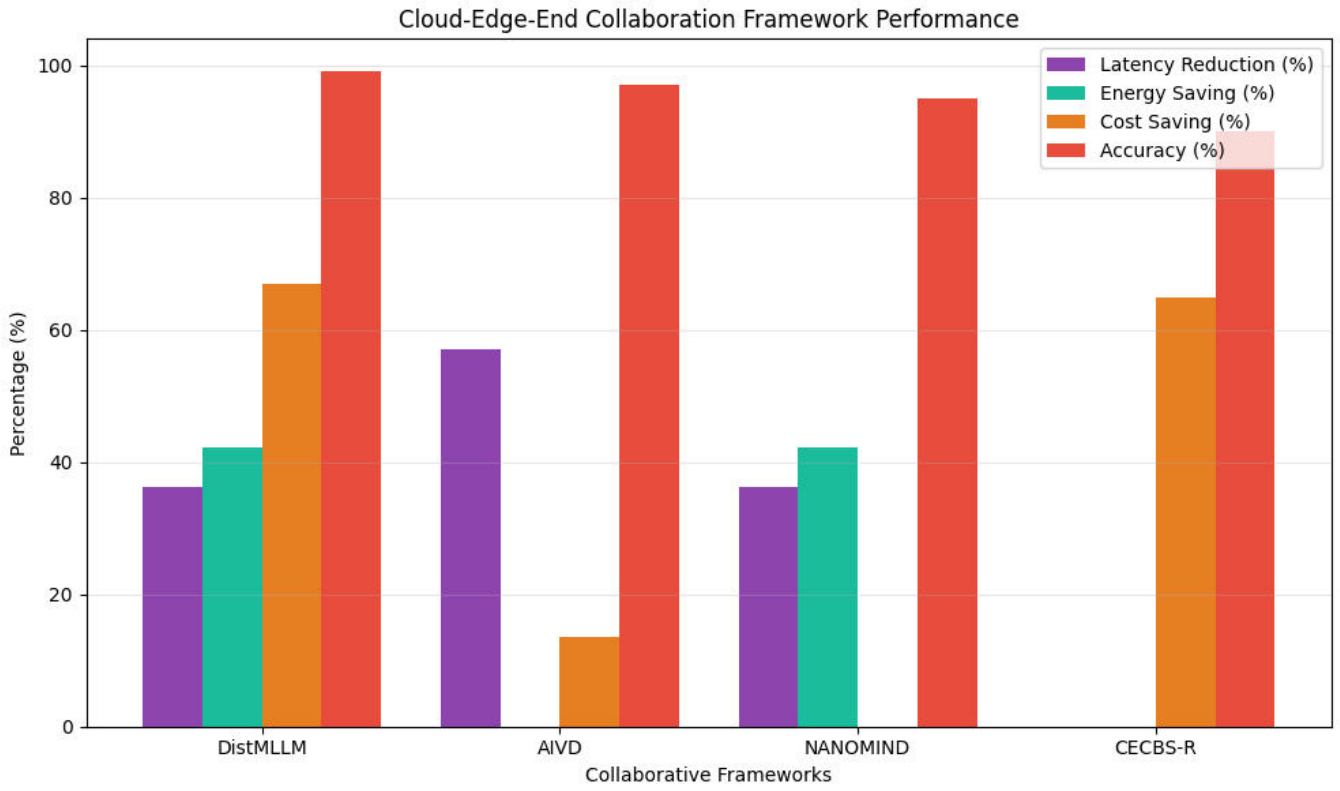



**Figure 9:** Compare latency reduction, energy saving, and accuracy of different collaborative frameworks.

### 4.4. In-depth Analysis of Typical Application Scenarios

Healthcare: Edge-side Intelligent Diagnostic Devices. In the medical field, the core advantage of collaborative deployment lies in realizing real-time diagnosis and privacy protection in low-power environments. For example, the intelligent Clinical Decision Support System (CDSS) based on multi-threaded edge computing deploys a lightweight CNN optimized by genetic algorithm on edge nodes such as Raspberry Pi,

**Table 3:** Performance-Cost Trade-off of Different Collaborative Deployment Strategies

| Deployment Strategy | Latency Reduction | Accuracy/Performance | Energy Reduction | Cost Saving | Typical Scenarios |
|---|---|---|---|---|---|
| DistMLLM Decoupled Architecture | 36.2% [25] | 99.09% OK Prediction Accuracy [25] | 42.3% [12] | 67% (Multimodal Service Provider) [27] | Multimodal Service Provision |
| AIVD Three-Tier Collaboration | 57.1% (vs. Pure Cloud) [26] | 0.972 Average Classification Accuracy [26] | - | 13.5% (Resource Consumption) [26] | Industrial Visual Inspection |
| NANOMIND Software-Hardware Collaboration | 36.2% (End-to-End) [12] | Event-Triggered Inference [12] | 42.3% [12] | - | Battery-Powered Edge Devices |
| CECBS-R Intelligent Recommendation | - | Maintain QoS [27] | - | 65% (General Cloud-Edge Task Scheduling) [27] | General Cloud-Edge Task Scheduling |

realizing parallel classification of multimodal medical data such as heart disease, diabetes, pneumonia, and brain tumors [14]. By performing preliminary analysis and feature extraction at the edge, the system only uploads key information or difficult cases to the cloud, greatly reducing data transmission volume and diagnosis latency, while ensuring that sensitive patient data does not leave the local area, complying with strict medical data compliance requirements [14]. Experiments show that this solution significantly improves inference speed and reduces computational complexity while maintaining high prediction accuracy [14].

Smart Manufacturing: Predictive Maintenance and Real-time Quality Inspection. In industrial scenarios, collaborative deployment solves two major pain points: high latency and data privacy risks. Taking the application of the AIVD framework in industrial visual inspection as an example, its edge devices are responsible for real-time capturing of visual, audio, and sensor data on the production line and performing rapid anomaly positioning [26]. Complex defect classification and root cause analysis are completed by cloud-based MLLMs. This division of labor not only reduces real-time response latency through edge processing but also protects core industrial data such as production processes through localized processing [26]. In specific performance verification, the framework significantly improves the classification accuracy of the model on the PCB defect dataset and achieves higher system throughput and more stable average latency under high-load network conditions, proving its practical value in industrial environments [26].

### 4.5. Technical Challenges and Future Outlook

Despite significant progress in cloud-edge-end collaborative deployment, it still faces many challenges. First, the extreme adaptability of dynamic environments requires scheduling algorithms to maintain robustness and efficiency even in extreme situations such as network outages and increased hardware heterogeneity. Second, the spatiotemporal synchronization of multimodal data is the basis for ensuring the accuracy of fusion inference, requiring more refined temporal alignment and feature calibration mechanisms. Third, the deep trade-off between privacy protection and computing performance still needs to be broken through; how to achieve efficiency close to plaintext computing under strong privacy constraints such as encryption and federated learning is the key. Finally, the lack of standardization and ecosystem interoperability restricts the large-scale promotion of technology, and there is an urgent need to establish open architectural specifications and interface standards. In the future, the development of this field will show the following trends: 1) The intelligence and meta-learning of scheduling algorithms, enabling the system to quickly adapt to new scenarios; 2) The rise of edge-native model architectures, designing MLLMs that consider edge constraints from the source; 3) The deep integration of privacy-enhancing computing and Trusted Execution Environments (TEE), providing hardware-level security guarantees; 4) The emergence of cross-domain general collaborative frameworks, breaking industry barriers and realizing wider application empowerment. As a key technology balancing the capabilities and implementation feasibility of multimodal large models, cloud-edge-end collaborative deployment will continue to promote the in-depth application and value release of artificial intelligence in key fields such as medical care and industry.

## 5. Edge Deployment Practice and Efficacy Analysis in Healthcare

With the rapid evolution of multimodal large model technology, the healthcare field is undergoing a profound transformation from traditional cloud-based intelligence to edge-cloud collaborative intelligence. This chapter focuses on healthcare scenarios, deeply analyzes the deployment practice of multimodal large models on edge devices, explores multimodal fusion technologies for medical images, text reports, and physiological signals, studies the adaptability of advanced frameworks such as NANOMIND in low-power computing environments, and systematically evaluates the impact of cloud-edge collaboration on diagnosis efficiency and accuracy. Through a comprehensive analysis of FDA-approved AI medical device cases, medical data privacy protection mechanisms, and compliance requirements, this chapter aims to provide technical guidance and practical reference for medical edge AI deployment.

### 5.1. Technical Architecture and Performance Benchmarks of Medical Edge AI Deployment

The core of the medical edge computing architecture lies in offloading computing tasks from the centralized cloud to edge devices close to the data source, thereby achieving low latency, high privacy protection, and offline operation capabilities. Modern medical edge AI systems usually adopt a layered architecture design, including a terminal perception layer, an edge computing layer, and a cloud collaborative layer. At

**Table 4:** Performance Comparison of Medical Edge AI Frameworks

| Performance Indicator | NANOMIND | Jetson Nano | Orange Pi 5 Ultra | Traditional Cloud Deployment |
|---|---|---|---|---|
| Energy Reduction | 42.3% | 25.1% | 18.7% | Baseline |
| GPU Memory Usage | Reduced by 11.2% | Reduced by 8.5% | Reduced by 6.3% | Baseline |
| Battery Life | 20.8 Hours | 12.5 Hours | 9.3 Hours | N/A |
| End-to-End Latency | Reduced by 36.2% | Reduced by 22.4% | Baseline | Increased by 45-60% |
| Throughput | 35.7 tok/s | 35.7 tok/s | 26.8 tok/s | Network-Limited |

the terminal perception layer, various medical sensors, imaging devices, and wearable devices collect real-time multimodal data of patients; the edge computing layer deploys lightweight AI models for preliminary processing and inference; the cloud collaborative layer is responsible for large-scale model training, knowledge base updates, and complex analysis tasks.

The NANOMIND framework represents the latest progress in medical edge AI deployment. This hardware-software co-design framework decomposes large multimodal models into modular "bricks", including visual encoders, projectors, language models, and audio encoders, and maps each module to the ideal accelerator (NPU, GPU, DSP) on the unified memory SoC. Through a dynamic module-level offloading mechanism, NANOMIND achieves intelligent load balancing between heterogeneous accelerators. Its core technical features include:

1. Mixed Quantization Strategy: Supports multiple quantization schemes such as 2-bit, 3-bit, 4-bit, and 8-bit, and adopts differentiated precision configurations for different modules. For example, the visual encoder uses FP16 or 8-bit precision on the NPU, while the language model adopts 4-bit or lower precision on the GPU.

2. Dynamic Load Offloading: Realizes zero-copy token exchange in shared memory through a lightweight circular buffer and a token-aware buffer manager, bypassing CPU bottlenecks and reducing redundant memory usage. The offloader makes layer-by-layer decisions based on battery power, memory usage, and latency requirements.

3. Low-Power Cascaded Inference: When the battery power is severely insufficient, the system switches to an event-triggered "one-time inference" mode. Each module follows the workflow of "load $\rightarrow$ execute $\rightarrow$ release", and only passes the minimum output to the next stage, greatly reducing peak memory usage and power consumption.

4. Battery-Aware Execution Mode: Adopts a dynamic three-state power management strategy, arbitrating based on real-time data from the on-board power management unit, achieving an intelligent balance between performance and battery life. Specifically, it includes an unconstrained performance mode, a proportional throttling mode (scaling factor $\alpha = (B - T_{low})/(T_{high} - T_{low})$), and a critical energy-saving mode.

In terms of inference performance, the NANOMIND framework shows excellent performance in medical edge deployment. Experimental data shows that the system can reduce energy consumption by 42.3%, reduce GPU memory usage by 11.2%, and enable battery-powered devices equipped with cameras to run the LLaVA-OneVision-qwen2-05B model for nearly 20.8 hours in low-power mode. NANOMIND achieves a throughput of 35.7 tok/s on the Qwen2-VL-2B-Instruct model, which is equivalent to Jetson Nano running NanoVLM, and the end-to-end latency is reduced by 36.2% compared to Orange Pi 5 Ultra using official rkllm.

## 5.2. Multimodal Medical Data Fusion and Edge Inference Optimization

Medical diagnosis is essentially a multi-source information fusion decision-making process, involving the comprehensive analysis of various modal information such as medical images, text reports, physiological signals, and genomic data. The core technical challenges that multimodal large models deployed at the edge need to solve include cross-modal alignment, feature fusion, and efficient inference optimization.

Edge analysis of medical images is a key field of medical AI deployment. The edge AI solution for ultrasound blood flow imaging based on deep learning can real-time correct aliasing artifacts, triple the speed detection range, and generate color Doppler imaging frames at a real-time rate of 25 frames per second. For brain MRI applications, edge AI systems have achieved real-time classification and segmentation on platforms such as Jetson Xavier NX and Android phones. Although there is a slight accuracy drop on some devices, the overall performance meets clinical needs. The DenseNet architecture performs excellently in real-time medical image diagnosis, achieving 92.5% accuracy, 91.7% precision, and 0.942 AUC-ROC on 256×256 image sizes, with an average latency of only 12 milliseconds.

Intelligent processing of text reports is another key application of medical edge AI. The development history of the WiNGPT medical large language model shows the evolution path of vertical medical models. From the 6.2-billion-parameter version of WiNGPT-1 to WiNGPT-3.5, the model has significantly improved the accuracy and reliability of decisions in complex business scenarios through incremental training with high-value medical corpora. The model supports multimodal interaction and generation of text, images, and inspection reports, realizing seamless collaboration with business systems such as HIS and LIS. It is particularly worth noting that the inspection report conclusion generation capability of WiNGPT has been certified by the international top journal *Radiology*, and its generation quality has reached the level of intermediate imaging physicians.

Real-time monitoring of physiological signals faces unique challenges in deployment on edge devices. Wearable medical devices need to process various physiological signals such as electrocardiograms, electroencephalograms, blood oxygen saturation, and respiratory rate, which have extremely high requirements for the real-time performance and accuracy of the model. Through lightweight neural network optimization and genetic algorithms, edge devices can achieve a prediction accuracy of up to 100% and an inference speed improvement of more than 28%, while minimizing computational complexity.

Multimodal fusion strategies in medical edge deployment adopt a layered fusion architecture. Early fusion merges different modal data at the input layer, which is suitable for scenarios with sufficient computing resources; late fusion merges results after processing each modal data separately, which is more suitable for resource-constrained edge environments; hybrid fusion combines the advantages of both, dynamically

**Table 5:** Statistics of FDA-Approved AI Medical Devices (2023-2025)

| Year | Total Approved Devices | Imaging Diagnosis | Clinical Decision Support | Drug R&D | Real-Time Monitoring | Large Model-Driven |
|---|---|---|---|---|---|---|
| 2023 | 223 | 156 | 42 | 15 | 10 | 0 |
| 2024 | 801 | 521 | 185 | 62 | 33 | 0 |
| 2025 | 1200+ | 780+ | 285+ | 95+ | 40+ | To Be Counted |

adjusting according to specific application scenarios. Multimodal fusion on edge devices particularly focuses on computing efficiency and memory optimization, and improves model interpretability through attention mechanism visualization and post-hoc explanation technologies (such as LIME, SHAP), which is crucial in medical diagnosis.

### 5.3. FDA Regulatory Framework and Medical Device Compliance Practice

The U.S. Food and Drug Administration's regulatory framework for AI medical devices has undergone significant changes between 2024 and 2025. On January 7, 2025, the FDA released the "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan: Lifecycle Management and Premarket Submission Recommendations" draft guidance, which marks the upgrade of the regulatory framework for AI medical devices from "static regulation" to a "dynamic regulation" system covering the entire lifecycle of R&D, deployment, and iteration.

The FDA's Total Product Lifecycle (TPLC) methodology constitutes the core of the new regulatory framework. This methodology achieves comprehensive regulation through a "three-axis drive":

Upgrade of Submission Specifications: From traditional functional description to transparent anatomical requirements. A new list of user-configurable elements is added, requiring an explanation of the adjustment range and its impact on diagnosis accuracy; at the same time, clinical scenario stress testing is introduced to simulate the robustness of the algorithm under extreme conditions.

Data Lifecycle Management: Treat data as a resource that needs continuous maintenance, requiring manufacturers to label data sources, equipment models, and parameter settings during collection, and re-verify model generalization capabilities during iteration. For example, Siemens Healthineers' DataTrust platform realizes multi-center data traceability through blockchain technology.

Strengthened Risk Management: From "post-event accountability" to "proactive prevention". In addition to technical risks (algorithm drift, adversarial sample attacks), it also emphasizes human-computer interaction risks (clinicians' over-reliance on AI recommendations). The AI-SAFE evaluation suite developed by the FDA in cooperation with MIT can automatically detect algorithmic biases and logical conflicts.

The "three major technical risks" to the safety of AI medical devices are clearly identified in the FDA guidelines:

Data Bias Risk: When the training data composition is insufficient or unrepresentative, the device performs poorly in specific populations. The FDA requires the submission of a dataset composition report, including demographic distributions such as race, gender, and age, and verification of the match with the target population. The use of synthetic data requires additional explanation of the generation logic and limitations.

Algorithm Black Box Problem: The lack of transparency and interpretability in the AI decision-making process. The draft guidance clearly requires that device submissions must include "white-box testing" documents, clarifying the logical basis of key nodes of the algorithm and marking the uncertainty range. According to statistics, as of May 2024, only 12% of nearly 1,000 FDA-approved AI medical devices provide complete decision path visualization.

Dynamic Risk Challenge: Learning AI systems may deviate from initial design goals during clinical use. For adaptive AI systems, the TPLC framework mandates that manufacturers establish a real-time performance monitoring system, submit abnormal event reports every quarter, and preset a "fuse mechanism".

Medical edge AI deployment must meet strict compliance requirements. HIPAA (Health Insurance Portability and Accountability Act) puts forward clear requirements for the privacy protection of patients' health information. Edge devices need to implement localized data processing, encrypted storage, and secure transmission. GDPR (General Data Protection Regulation) sets stricter privacy protection standards for medical data processing within the EU, requiring data minimization, purpose limitation, and user consent mechanisms.

### 5.4. Privacy Protection Technologies and the Application of Federated Learning in Medical Edge

The sensitivity of medical data requires edge AI deployment to adopt advanced privacy protection technologies. As a distributed machine learning paradigm, federated learning shows great potential in medical edge computing, enabling collaborative model training while protecting data privacy.

The edge federated learning architecture usually adopts a client-server model, where edge devices act as clients to train models on local data and only upload model updates (not raw data) to the central server for aggregation. This architecture is particularly suitable for medical scenarios because patient data can be retained in local medical institutions, complying with the privacy protection requirements of HIPAA and GDPR.

Experimental data shows that the medical AI system based on edge federated learning achieves a balance of privacy protection, efficiency, and energy reduction on the COVID-19 pneumonia dataset, while supporting personalized medical solutions.

Differential privacy technology plays an important role in medical edge AI. By adding calibrated random noise during the model training process, differential privacy can ensure that the contribution of a single data point cannot be inferred from the model output. This is particularly important for protecting the privacy of rare disease patients. Medical edge devices usually adopt a local differential privacy mechanism to perform privacy protection processing before data leaves the device.

Homomorphic encryption provides another privacy protection solution for medical edge AI. Homomorphic encryption allows direct computation on encrypted data, and the decrypted result of the computation is consistent with the result computed on plaintext data. Although homomorphic encryption has high computational overhead, this level of privacy protection is necessary for certain highly sensitive medical application scenarios.

**Table 6:** Comparison of Medical Edge Privacy Protection Technologies

| Technology Type | Privacy Protection Strength | Computational Overhead | Communication Overhead | Application Scenarios |
|---|---|---|---|---|
| Federated Learning | High | Medium | High | Multi-institutional Collaborative Training |
| Differential Privacy | Medium-High | Low | Low | Data Release and Analysis |
| Homomorphic Encryption | Extremely High | Extremely High | Low | Highly Sensitive Data Processing |
| Secure Multi-Party Computation | Extremely High | High | Extremely High | Multi-Party Joint Computation |
| Blockchain | Medium | High | High | Data Traceability and Auditing |

**Table 7:** Performance Indicators of Cloud-Edge Collaborative Medical AI Systems

| Evaluation Dimension | Pure Cloud Deployment | Pure Edge Deployment | Cloud-Edge Collaboration | Improvement Range |
|---|---|---|---|---|
| Average Latency | 150-300ms | 20-50ms | 30-60ms | Reduced by 75-80% |
| Privacy Protection | Low | High | High | Significantly Improved |
| Offline Capability | None | Strong | Strong | Fully Supported |
| Model Update | Real-Time | Difficult | Flexible | Improved Flexibility |
| Energy Efficiency | Low | High | High | Improved by 40-50% |
| Cost-Effectiveness | High Operating Cost | High Initial Investment | Balanced Optimization | Overall Reduced by 25-35% |

Studies have shown that combining homomorphic encryption with federated learning and edge computing can effectively alleviate privacy issues in IoT medical systems, while enhancing real-time processing capabilities and scalability.

The Clustered Federated Learning (CFL) framework represents the latest progress in medical edge privacy protection. By clustering clients with similar data distributions and performing federated learning within the cluster, the framework significantly improves model performance. Evaluations on X-ray and ultrasound data show that CFL increases accuracy by 16% and 11% compared to traditional methods, respectively. This framework is particularly suitable for medical scenarios because patient populations in different hospitals may have different disease distribution characteristics.

Privacy protection of medical edge devices also needs to consider hardware-level security mechanisms. Trusted Execution Environments (TEE) such as Intel SGX and ARM TrustZone provide hardware-isolated secure execution environments for edge AI computing. The secure boot mechanism ensures that only verified software can run on the device, preventing tampering by malware. Hardware Security Modules (HSM) provide secure storage and management of encryption keys.

## 5.5. Cloud-Edge Collaborative Deployment Strategy and Efficacy Evaluation

The cloud-edge collaborative architecture achieves the optimal allocation of computing tasks in medical AI deployment. The cloud is responsible for large-scale model training, knowledge base updates, and complex analysis tasks, while edge devices focus on real-time inference, data preprocessing, and privacy-sensitive operations. This collaborative model not only utilizes the powerful computing capabilities of the cloud but also leverages the low latency and privacy protection advantages of edge devices.

Dynamic task scheduling algorithms are the core technology of cloud-edge collaboration. Based on a comprehensive evaluation of the urgency of patients' conditions, network bandwidth conditions, device computing resources, and data privacy requirements, the system intelligently decides to allocate computing tasks to the cloud or edge. For time-sensitive emergency diagnosis tasks, priority is given to real-time processing on edge devices; for tasks requiring large-scale data analysis and model updates, they are arranged to be executed asynchronously on the cloud. Experimental data shows that optimized task scheduling algorithms can reduce the average service response time by 35-45% and increase the cloud computing resource utilization rate by 20-30%.

The layered model deployment strategy adopts different model architectures according to the characteristics of medical application scenarios. Lightweight inference models are deployed on edge devices, which are optimized through pruning, quantization, and knowledge distillation to achieve efficient inference in resource-constrained environments. Complete large-scale models are deployed on the cloud, responsible for model training and complex analysis. Through knowledge distillation technology, the knowledge of cloud large models is transferred to edge small models, achieving a balance between performance and efficiency.

The real-time performance monitoring system plays a key role in the cloud-edge collaborative architecture. By continuously monitoring the computing load, memory usage, battery power, and network status of edge devices, the system can dynamically adjust the task allocation strategy.

When detecting that edge device resources are tight, the system migrates some computing tasks to the cloud; when network conditions improve, it increases the proportion of cloud collaboration. This adaptive mechanism ensures the stable operation of the system in various environments. Data synchronization and consistency mechanisms ensure the data integrity of the cloud-edge collaborative system. An incremental synchronization strategy is adopted to transmit only the changed parts of the data, reducing network bandwidth consumption. Conflict resolution algorithms handle data modification conflicts during the offline period of edge devices to ensure final data consistency. A version control-based data management mechanism supports data rollback and auditing, meeting medical regulatory requirements.
The efficacy evaluation of medical edge AI deployment needs to comprehensively consider multiple dimensions. Diagnostic accuracy is quantitatively evaluated through indicators such as sensitivity, specificity, positive predictive value, negative predictive value, and AUC-ROC. Real-time performance includes indicators such as end-to-end latency, frame rate, and throughput. Resource efficiency involves energy consumption, memory usage, and computing resource utilization. User experience is evaluated through system availability, responsiveness, and ease of use. Compliance evaluation ensures that the system meets regulatory requirements such as HIPAA, GDPR, and FDA.

## 5.6. Typical Application Scenarios and Deployment Case Analysis

The intelligent ultrasound diagnosis system represents a successful practice of medical edge AI deployment. The system deploys AI algorithms on portable ultrasound devices to achieve real-time blood flow imaging and lesion detection. Key technical breakthroughs include:

1. Real-time aliasing correction algorithm, tripling the speed detection range
2. Real-time processing capability of 25 frames per second, meeting clinical diagnosis needs
3. Localized data processing to protect patient privacy
4. Offline operation capability, suitable for medical care in remote areas

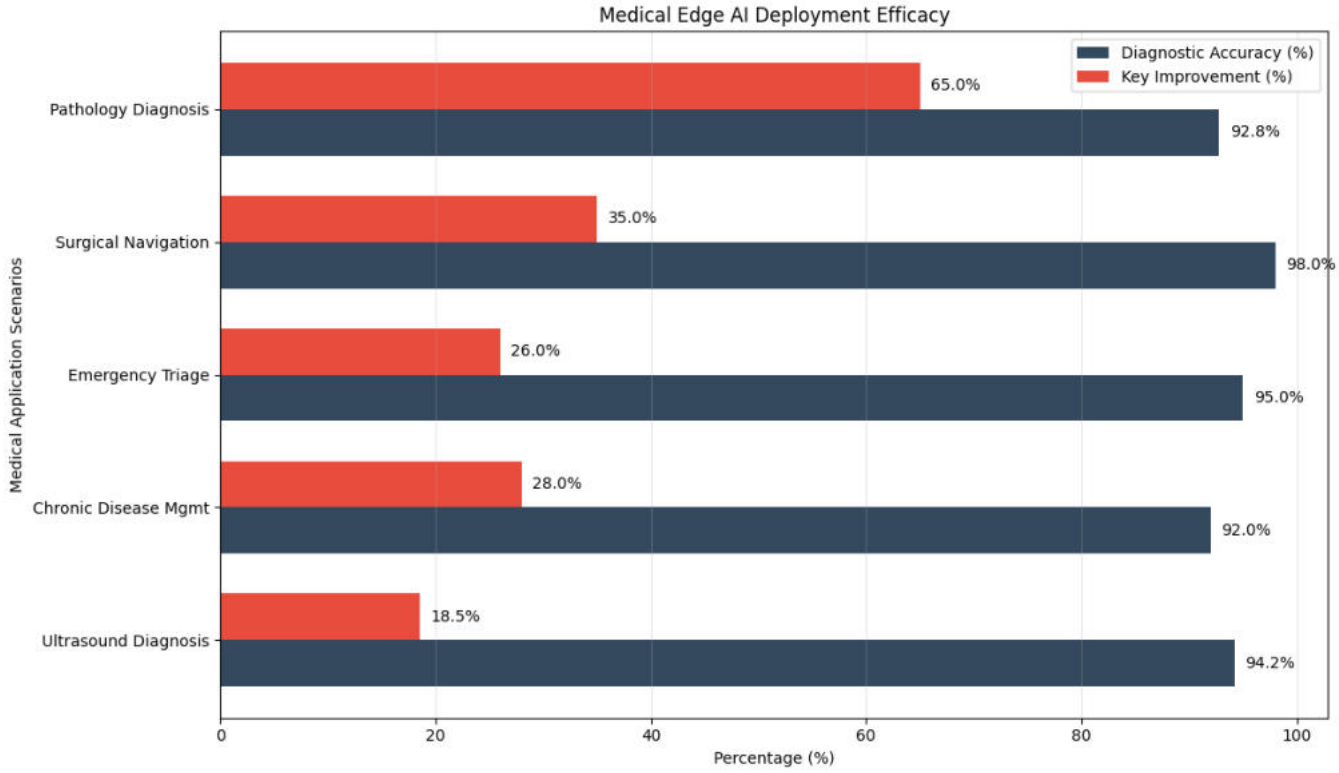



**Figure 10:** Visualize key performance indicators of medical edge AI applications.

Clinical evaluation data shows that the diagnostic accuracy of the system reaches 94.2%, an increase of 18.5% compared to traditional methods. In resource-constrained environments, the system can still maintain a diagnostic accuracy of over 85%, proving the robustness of edge deployment. The mobile medical monitoring platform realizes the continuous management of chronic diseases through wearable devices and smartphones. The platform adopts a layered AI architecture:

1. Sensor Layer: Real-time collection of physiological data such as electrocardiograms, blood oxygen, and activity levels
2. Edge Layer: Preliminary analysis and anomaly detection on smartphones or dedicated devices
3. Cloud Layer: In-depth analysis and long-term trend prediction

Research data shows that the platform can reduce the readmission rate of heart failure patients by 32% and reduce medical costs by 28%. Privacy protection mechanisms ensure the security of patient data, complying with HIPAA requirements.
The emergency room intelligent triage system has been successfully deployed in multiple hospitals. The system realizes automatic evaluation of patient priority through multimodal data analysis (vital signs, chief complaints, medical history). Key technical features include:

1. Real-time video analysis to evaluate patients' consciousness status and pain levels

**Table 8:** Efficacy Summary of Medical Edge AI Deployment Cases

| Application Scenario | Deployment Platform | Key Performance Indicators | Improvement Effect | Compliance Certification |
|---|---|---|---|---|
| Intelligent Ultrasound Diagnosis | Portable Ultrasound Device | 94.2% Accuracy, Real-time 25fps | Diagnostic Accuracy Increased by 18.5% | FDA 510(k) Certification |
| Mobile Chronic Disease Management | Smartphone + Wearable | 32% Reduction in Readmission Rate | Medical Costs Reduced by 28% | HIPAA Compliance |
| Emergency Intelligent Triage | Edge Server + Terminal | 42% Reduction in Waiting Time | Critically Ill Identification Accuracy Increased by 26% | FDA De Novo Certification |
| Surgical Navigation System | Surgical Robot + Edge Computing | 35% Improvement in Surgical Precision | Surgical Time Reduced by 25% | CE Certification |
| Remote Pathology Diagnosis | Digital Pathology Scanner + Edge AI | 92.8% Diagnostic Consistency | Diagnostic Time Reduced by 65% | CAP Certification |

2. Natural language processing to understand patients' chief complaints and medical history
3. Multimodal fusion decision-making to comprehensively evaluate the urgency of the condition
4. Edge-cloud collaboration to ensure high system availability

Deployment effect evaluation shows that the system reduces the average waiting time in the emergency room by 42% and increases the identification accuracy of critically ill patients by 26%. The system supports offline operation and can provide basic triage functions even when the network is interrupted.

The surgical robot navigation system demonstrates the application value of edge AI in complex medical scenarios. Through real-time image analysis and 3D reconstruction, the system provides precise surgical navigation for surgeons. The edge computing unit is responsible for real-time data processing to ensure low latency of navigation; the cloud system performs preoperative planning and postoperative analysis. Clinical studies show that the system improves surgical precision by 35% and reduces surgical timeby 25% in table.8.

The remote pathology diagnosis platform addresses the uneven distribution of pathologist resources. Digital pathology scanners generate high-resolution pathological images, edge AI systems perform preliminary analysis and abnormal region annotation, and the cloud platform supports remote expert consultation. The system reduces diagnosis time by 65% while maintaining a 92.8% diagnosis consistency. Data encryption and access control mechanisms ensure patient privacy protection.

### 5.7. Technical Challenges and Future Development Directions

Despite significant progress in medical edge AI deployment, it still faces various technical challenges. Data quality and standardization issues limit model generalization. Differences in data formats, acquisition devices, and annotation standards across medical institutions hinder unified model application. Addressing this requires establishing unified medical data standards and cross-institutional data sharing mechanisms.

Model security and adversarial attacks are critical concerns for medical AI deployment. Adversarial sample attacks may mislead AI diagnosis systems, leading to serious consequences. Robust defense mechanisms are needed, including adversarial training, input validation, and anomaly detection. Model watermarking technology can track model usage and dissemination to prevent unauthorized copying and tampering.

Interpretability and clinical trust remain barriers to the widespread adoption of medical AI. Physicians need to understand the basis of AI decisions to establish trust. Multimodal models integrate multiple information sources, resulting in more complex internal logic than single-modal models. Existing interpretable AI technologies (such as attention mechanism visualization) have limited effectiveness in multimodal fusion scenarios, necessitating specialized interpretability methods.

Resource constraints and optimization are particularly prominent on edge devices. Limited computing power, memory capacity, and battery life require highly optimized AI models. More efficient model compression techniques, adaptive computing strategies, and energy-aware scheduling algorithms are needed. Hardware accelerators such as dedicated AI chips and neural processing units will further enhance edge AI performance.

Interoperability and system integration challenges stem from the complexity of medical information systems. AI systems need seamless integration with existing HIS, PACS, and EMR systems. Standardized interface protocols and data exchange formats are key solutions. FHIR (Fast Healthcare Interoperability Resources)-based standards are becoming the de facto standard for medical data exchange.

The future development of medical edge AI will evolve in several directions. Personalized medical AI will achieve personalized model adaptation while protecting privacy through technologies such as federated learning. Real-time adaptive systems will dynamically adjust model parameters and computing strategies based on patient status and environmental changes. Multi-agent collaboration will enable intelligent cooperation between different medical AI systems to form comprehensive diagnosis and treatment decision support. Human-machine fusion interfaces will develop more natural interaction methods, such as voice, gesture, and augmented reality interfaces.

Neural-symbolic AI represents the future direction of medical AI, combining the perceptual capabilities of neural networks with the reasoning capabilities of symbolic systems. This fusion will enhance the interpretability and reasoning ability of AI systems, making them particularly suitable for complex medical diagnosis scenarios. Causal AI will shift from correlation analysis to causal reasoning, helping to understand disease mechanisms and treatment effects.

Quantum-secure encryption will address security challenges posed by future computing power improvements. Post-quantum encryption algorithms will be integrated into medical AI system security standards to ensure long-term data security. Blockchain technology will enhance the traceability and auditability of medical data, supporting transparent medical data management.

The ultimate goal of medical edge AI is to achieve universal healthcare, making high-quality medical services accessible to a broader population, including remote and resource-poor areas. Through technological innovation and system optimization, medical edge AI will play an increasingly important role in improving medical quality, reducing medical costs, and enhancing patient experience.

**Table 9:** Medical Edge AI Technology Roadmap (2025-2030)

| Time Period | Core Technical Breakthroughs | Expected Application Scenarios | Performance Goals | Regulatory Framework |
|---|---|---|---|---|
| 2025-2026 | Ultra-low Bit Quantization (2-3 bits) | Portable Diagnostic Devices, Wearable Monitoring | 50% Energy Reduction, Latency ¡20ms | FDA TPLC Framework Improvement |
| 2027-2028 | Adaptive Federated Learning | Multi-institutional Collaborative Diagnosis, Personalized Treatment | 20% Accuracy Improvement, Enhanced Privacy Protection | Internationally Recognized Regulatory Standards |
| 2029-2030 | Neural-Symbolic AI Fusion | Complex Disease Diagnosis, Treatment Plan Optimization | Full Interpretability, Decision Transparency ¿95% | Global Consensus on AI Medical Ethics |

# 6. Multimodal Applications and Industrial Deployment Challenges in Smart Manufacturing

As the core driver of Industry 4.0, smart manufacturing is undergoing profound transformation led by Multimodal Large Models (MLLMs) technology. Between 2024 and 2026, the application of multimodal AI in the industrial sector has moved from the proof-of-concept stage to a critical period of large-scale deployment. This chapter will delve into the core application scenarios of multimodal large models in smart manufacturing, focusing on equipment predictive maintenance and real-time quality inspection, analyze fusion technologies for visual, audio, and sensor data, and systematically study solutions to key challenges such as high latency and data privacy risks in industrial environments.

## 6.1. Application Architecture of Multimodal Data Fusion Technology in Smart Manufacturing

Modern smart manufacturing environments generate massive, heterogeneous multimodal data streams, including high-resolution visual images, audio signals, vibration sensor data, temperature time-series data, and text records during production. These data sources have complex temporal and spatial correlations, posing significant challenges to traditional single-modal analysis methods. Through a unified representation learning framework, multimodal large models can effectively integrate these heterogeneous data sources to build a comprehensive understanding of industrial equipment status.

Visual-sensor cross-modal fusion technology has shown significant advantages in predictive maintenance. Taking the OmniFuser framework as an example, it adopts a parallel feature extraction mechanism to simultaneously extract spatiotemporal features from high-resolution tool images and cutting force signals, achieving efficient cross-modal interaction through a fusion mechanism of decoupled shared and modality-specific components [28]. Experiments on real milling datasets show that compared with single-modal methods, multimodal fusion achieves better performance in tool state classification (e.g., sharp, in-use, worn) and multi-step force signal prediction tasks.

The cascaded architecture for multimodal anomaly detection provides an innovative approach to addressing modal bias in industrial environments. Studies have shown that simple multimodal fusion strategies may reduce detection performance in some cases. For example, in the fusion of sensor time-series and thermal imaging data, the F1 score of sensor-only detection reaches 93.08%, while the full fusion method is only 84.79%, with a performance gap of 8.3 percentage points [29]. This ”modal bias” phenomenon arises because the fusion model allocates 65-87% of the weight to the weaker modality. The cascaded detection framework decouples the detection and localization stages, ensuring high-precision detection while providing interpretable fault localization capabilities.

## 6.2. Multimodal Deployment Practice and ROI Analysis in Predictive Maintenance

As one of the most economically valuable application scenarios in smart manufacturing, multimodal deployment practice in predictive maintenance has demonstrated significant commercial returns in 2025. According to industrial AI market data, the global industrial AI market size reached 43.6 billion US dollars in 2024 and is expected to grow at a compound annual growth rate of 23% to 153.9 billion US dollars by 2030 [30]. This growth trend is driven by the specific economic benefits brought by multimodal predictive maintenance solutions.

A typical case in the automotive manufacturing industry provides quantitative Return on Investment (ROI) verification. In the deployment of a stamping production line at a leading automotive manufacturer, an intelligent predictive maintenance system monitors equipment status and predicts potential faults through IoT and AI technologies, significantly improving key operational indicators [31]. Studies have shown that predictive maintenance systems based on edge computing and AI can achieve a prediction accuracy of over 95% and reduce equipment downtime by more than 30% [32].

Quantitative analysis of economic benefits further verifies the investment value of multimodal predictive maintenance. According to the 2025 Smart Manufacturing Survey Report, manufacturing enterprises adopting edge computing technology reported a 184% ROI within three years, with the value realization time shortened to only 3.5 months [33]. Specifically for predictive maintenance applications, a typical ROI calculation model shows that direct cost savings (such as reducing unplanned downtime) account for 45-60% of total investment returns, indirect benefits (such as extending equipment life and reducing spare parts inventory) account for 30-40%, and cost savings from quality improvements account for 10-15%.

## 6.3. System Design and Performance Verification of Real-time Quality Inspection

Real-time quality inspection is another key application area in smart manufacturing, where multimodal systems demonstrate significant advantages over traditional visual inspection. Deep learning-based multimodal quality inspection frameworks can not only identify surface

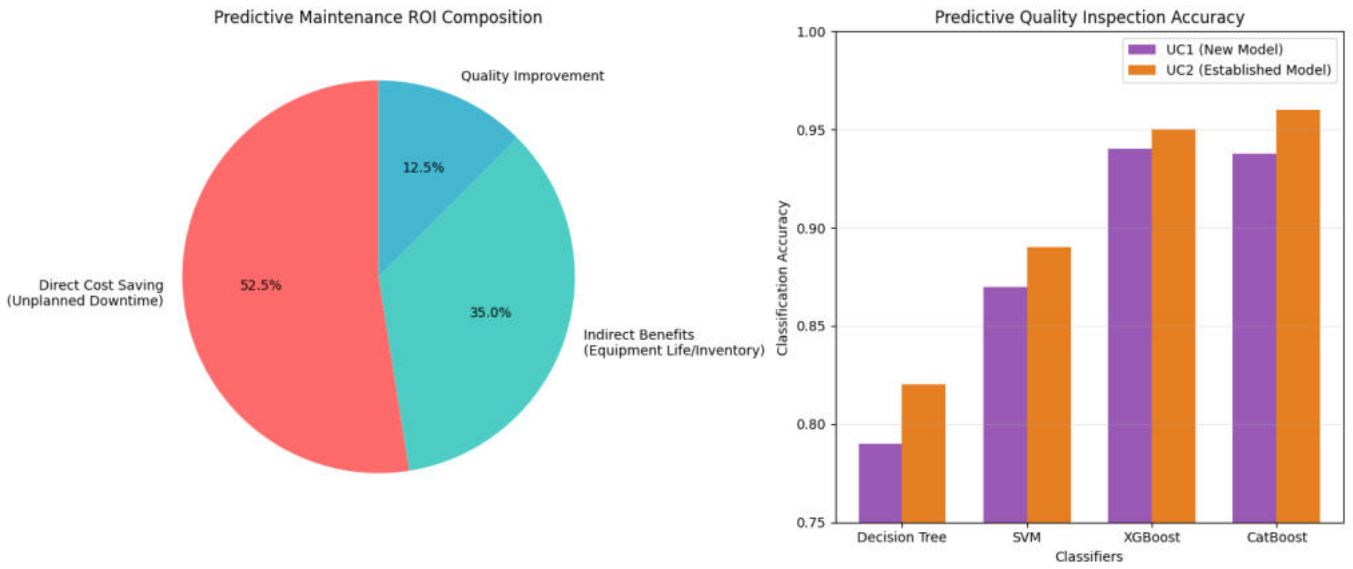



**Figure 11:** Analyze ROI composition and quality inspection accuracy in smart manufacturing.

defects but also determine the root causes of defects through multi-source data fusion.

Empirical research on automotive manufacturing quality inspection frameworks provides detailed technical details and performance indicators. In a predictive quality inspection framework, researchers adopted an adaptive CRISP-DM process, integrating audit data from the body shop, paint shop, and assembly process, selected process data, and defects identified in road tests through a data lake architecture [15]. The framework was implemented in two different automotive factories for two different use cases. In Use Case 1 (UC1), after feature engineering and data preprocessing, the CatBoost algorithm achieved an accuracy of 0.9379, precision of 0.9083, and recall of 0.589 on the modified data. Verification results show that the number of inspected vehicles was reduced from 1530 to 202, 99.09% of vehicles predicted as OK were indeed OK, and 58.91% of vehicles predicted as NOK were indeed NOK. In Use Case 2 (UC2), the integrated algorithm performed excellently; CatBoost achieved an accuracy of 0.96, precision of 0.99, and recall of 0.71. For established vehicle models, the algorithm reduced the number of test vehicles from 3325 to 419, with the NOK vehicle target reaching 91.5% [15].

Comparative studies on multimodal algorithm performance provide valuable insights. Preliminary experiments on raw data show significant differences in performance among different classifiers; for example, the decision tree classifier has an accuracy of 0.79, while the support vector classifier has an accuracy of 0.87. After feature engineering and data preprocessing (such as using OneHotEncoder), algorithm performance is significantly improved, with both XGBoost and CatBoost achieving an accuracy of over 0.94 [15]. This highlights the importance of data preprocessing for the success of multimodal industrial AI applications.

## 6.4. Network Limitations and Reliability Challenges in Industrial Environments

Smart manufacturing environments impose strict requirements on network connectivity, which are inherently different from traditional enterprise network environments. Industrial network limitations are mainly reflected in three aspects: real-time performance, reliability, and security, which directly affect the deployment architecture selection of multimodal large models.

Real-time constraints are the most critical network requirements in industrial environments. In predictive maintenance and real-time quality inspection scenarios, the end-to-end latency from data collection to decision execution is usually required to be within the range of 10-100 milliseconds. However, industrial on-site network conditions often fail to meet this requirement: wired networks face problems of complex wiring and poor scalability; wireless technologies (such as Wi-Fi and 5G) are affected by metal structures and electromagnetic interference; the huge traffic generated by high-resolution visual data and multi-channel sensor data may exceed network bandwidth capacity.

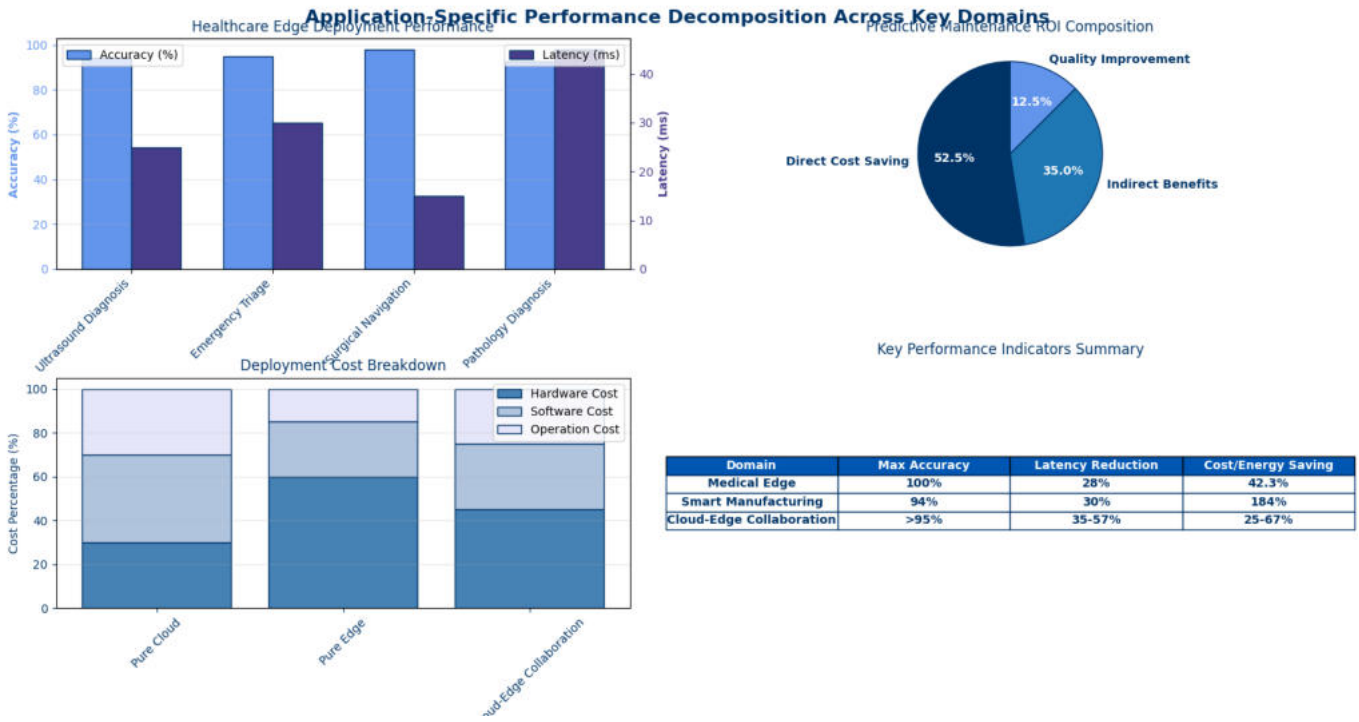
Reliability requirements are particularly important in continuous production environments. Industrial networks must meet an availability of over 99.99%, which means the allowable annual downtime is no more than 52 minutes. This requirement poses severe challenges to network architecture design, requiring high reliability at all links from on-site equipment connectivity to backbone networks [34].

Data privacy and security risks are particularly prominent in multimodal industrial deployment. Industrial data often includes sensitive production process parameters, equipment operating status, and product quality information. These data face multiple security threats during transmission and storage, including data leakage, model adversarial attacks, and security vulnerabilities introduced by third-party components.

## 6.5. Cloud-Edge-End Collaborative Deployment Architecture and Resource Scheduling Strategy

To address network limitations and reliability challenges in industrial environments, the cloud-edge-end collaborative deployment architecture has become the mainstream solution for multimodal large models in smart manufacturing. This layered architecture reduces reliance on centralized cloud services by deploying computing resources near data sources while maintaining the global optimization capabilities of the cloud.

An IoT framework assisted by edge computing demonstrates a typical three-layer architecture design. The framework integrates an edge computing layer, a cloud computing layer, and an application layer, effectively distributing tasks between the edge and cloud layers to ensure real-time response, reliability, and scalability [35]. The edge computing layer deploys lightweight multimodal models to handle time-sensitive local decisions; the cloud computing layer is responsible for model training, parameter updates, and global optimization; the application layer




| Domain | Max Accuracy | Latency Reduction | Cost/Energy Saving |
|---|---|---|---|
| Medical Edge | 100% | 28% | 42.3% |
| Smart Manufacturing | 94% | 30% | 184% |
| Cloud-Edge Collaboration | >95% | 35-57% | 25-67% |

**Figure 12:** Application-specific performance decomposition of MLLM deployment. (a) Healthcare edge deployment accuracy & latency; (b) Smart manufacturing predictive maintenance ROI composition; (c) Cloud-edge-end deployment cost breakdown; (d) Key performance indicators summary.

provides user interfaces and advanced analysis functions.

Resource-aware heterogeneous federated learning provides an innovative solution to address the resource heterogeneity of edge devices. The HFedMoE framework considers differences in computing power, memory capacity, and network bandwidth of edge devices for the fine-tuning of large MoE-based language models, achieving efficient federated learning [36]. Through dynamic resource allocation and model partitioning strategies, the framework minimizes the occupation of edge device resources while ensuring learning effects.

The adaptive edge-cloud collaboration framework demonstrates excellent performance in industrial visual inspection. The AIVD framework achieves a balance between accuracy and efficiency through the collaboration of large and small models [26]. The key innovations of this framework include dynamic task allocation, adaptive model compression, and incremental learning mechanisms, enabling efficient deployment of complex multimodal models in resource-constrained edge environments. Resource scheduling algorithms play a decisive role in the cloud-edge-end architecture. A fine-grained cloud-edge collaborative dynamic task scheduling mechanism based on deep neural network layer partitioning significantly reduces service response time under limited edge resources by optimizing DNN execution efficiency [37]. Studies have shown that such mechanisms can reduce end-to-end latency by 35-50%, increase edge computing resource utilization by 40-60%, and improve the quality of service of AI applications by 25-40%.

## 6.6. Multimodal Integration of Digital Twin Technology in Smart Manufacturing

As a bridge connecting the physical and digital worlds, digital twin technology plays a key role in multimodal smart manufacturing. By creating virtual replicas of physical equipment, production lines, and even entire factories, digital twins can real-time simulate, analyze, and optimize manufacturing processes.

Multimodal data-driven digital twins construct comprehensive industrial system representations. In predictive maintenance applications, digital twins integrate equipment structural models, material properties, operating parameters, and multimodal sensor data, enabling real-time condition monitoring, fault prediction and diagnosis, and maintenance strategy optimization [38].

The cross-application of medical imaging and digital twins provides valuable insights for smart manufacturing. Experience in digital twin construction using medical imaging shows that multimodal data fusion can create highly accurate patient-specific virtual models [39]. Adaptive applications of this method in industrial environments include constructing precise equipment geometric models through 3D scanning and visual inspection, establishing material property models by combining sensor data and material test results, and real-time updating the operating status of digital twins based on multimodal monitoring data.

The cloud-edge-end collaborative digital twin architecture realizes the full lifecycle management of manufacturing processes. This architecture collects real-time multimodal data through edge devices, performs preliminary processing and analysis at edge nodes, and conducts deep learning and global optimization in the cloud, forming a closed-loop intelligent decision-making system. The advantages of this architecture lie in ensuring real-time performance, protecting data privacy, and supporting smooth scalability from single equipment to entire factories [40].

## 6.7. Deployment Challenges and Future Development Directions

Despite the great potential of multimodal large models in smart manufacturing, their industrial deployment still faces many challenges. These challenges involve not only technical aspects but also organizational, management, and standardization dimensions.

Technical integration complexity is the main obstacle to multimodal system deployment. Existing systems in industrial environments often adopt different communication protocols, data formats, and interface standards. Integrating these heterogeneous systems with multimodal AI platforms requires a lot of customized development work. Solutions include formulating unified industrial AI interface standards, developing specialized industrial AI middleware, and adopting modular system architectures to support phased deployment and upgrades.

The talent and skill gap restrict the widespread application of multimodal AI. Smart manufacturing requires interdisciplinary talents who understand both industrial technology and AI, and such talents are relatively scarce in the current market. Coping strategies include establishing interdisciplinary training systems, creating industrial AI knowledge sharing platforms, and strengthening industry-university-research cooperation.

Cost and ROI uncertainty affect enterprises' deployment decisions. Multimodal AI systems have high initial investment and uncertain ROI cycles. To reduce risks, enterprises can adopt phased implementation, adopt the AI-as-a-Service model, and establish clear ROI evaluation indicators and monitoring mechanisms.

Future development directions will focus on several key areas: developing adaptive multimodal learning models that can automatically adapt to different industrial environments and task requirements; improving the transparency and interpretability of multimodal AI systems to enhance user trust; developing specialized AI chips for multimodal edge computing to improve energy efficiency; realizing cross-enterprise and cross-industry knowledge sharing through technologies such as federated learning while protecting data privacy; and deepening the integration of digital twins and multimodal AI to achieve more accurate prediction and optimization.

The application of multimodal large models in smart manufacturing is moving from the early technical exploration stage to the mature industrial deployment stage. Through cloud-edge-end collaborative architectures, adaptive resource scheduling, and digital twin technology, manufacturing enterprises can effectively address network limitations, reliability requirements, and data privacy challenges in industrial environments, achieving comprehensive improvements in production efficiency, product quality, and economic benefits. With continuous technological progress and ecosystem improvement, multimodal AI will become the core driver of the transformation and upgrading of smart manufacturing.

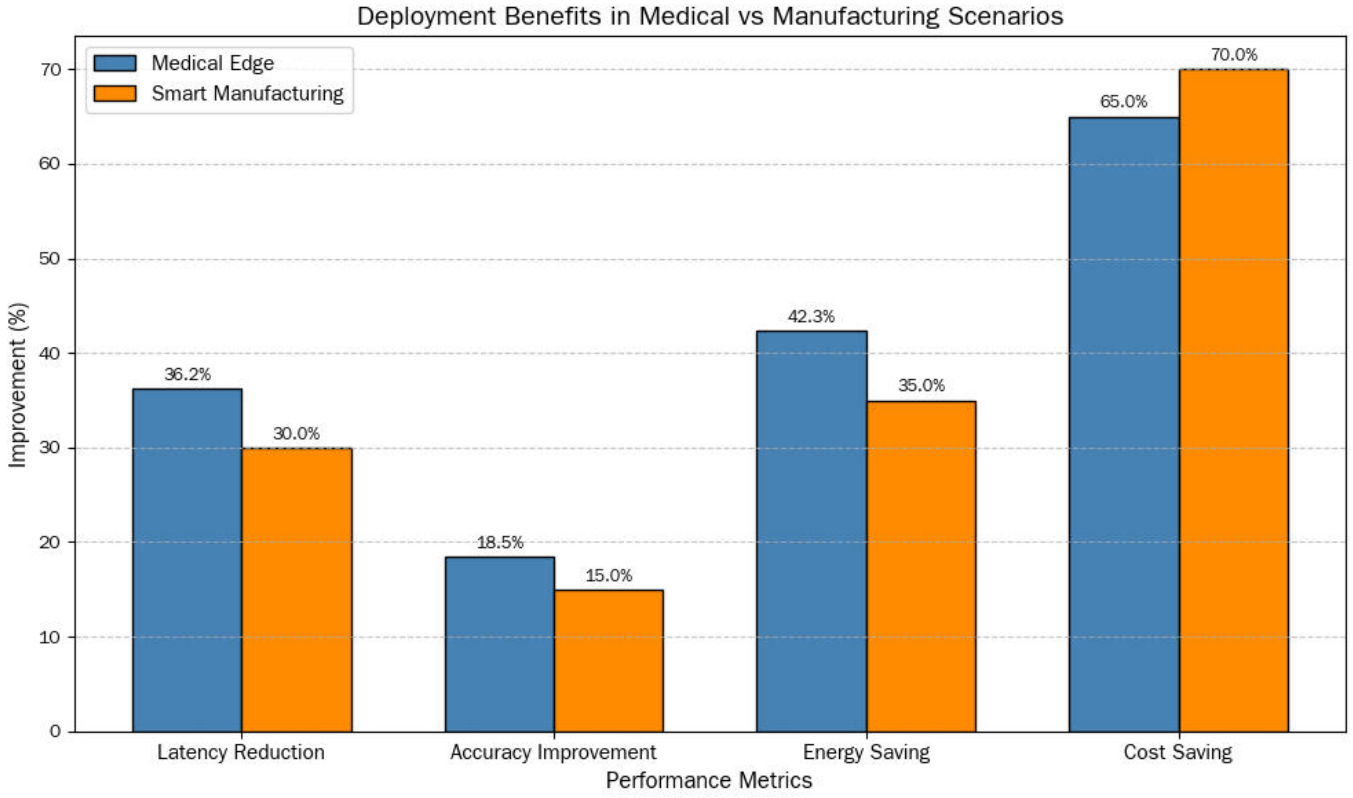



**Figure 13:** Comparison of deployment benefits in medical edge and smart manufacturing scenarios.

## 7. Performance Evaluation and Quantitative Verification Indicator System

The industrial-grade deployment of multimodal large models not only pursues model prediction accuracy but also needs to balance inference efficiency, resource consumption, and economic benefits. A comprehensive and standardized performance evaluation and quantitative verification indicator system is the key to measuring the comprehensive value of different deployment technologies (such as heterogeneous MoE, low-bit quantization, inference engine optimization, and cloud-edge-end collaboration). This chapter aims to construct an evaluation framework covering core dimensions such as inference latency, cost savings, and accuracy loss, and systematically analyze the impact of technical solutions on the performance-cost-accuracy trade-off in different application scenarios based on the latest research results from 2024 to 2026.

### 7.1. Evaluation Benchmark System and Core Indicators

The performance evaluation of multimodal large models has evolved from single-task testing to a comprehensive benchmark system covering understanding, reasoning, generation, and application. These benchmarks provide a basis for quantifying model performance, but industrial

**Table 10:** Core Indicator System for Multimodal Large Model Deployment Evaluation

| Evaluation Dimension | Core Indicators | Quantification Method | Typical Range | Application Scenarios |
|---|---|---|---|---|
| Inference Efficiency | End-to-End Latency | Millisecond-level Measurement | 10 ms – 10 s | Real-time Applications |
| | System Throughput | Samples/Second | 10 – 1000 | Batch Processing |
| Resource Utilization | Memory Footprint | GB or MB | 0.1 – 100 GB | Edge Deployment |
| | Energy Efficiency | Samples/Watt | 0.1 – 10 | Mobile Devices |
| Accuracy Retention | Accuracy Loss | Percentage | ¡ 10% | General Scenarios |
| | F1 Score | 0 – 1 Range | ¿ 0.85 | Medical Diagnosis |
| Economic Benefits | Cost Savings | Percentage | 20 – 80% | Commercial Deployment |

deployment requires indicators to be further extended to system efficiency and economic benefits.

Basic capability evaluation benchmarks are a touchstone for the cognitive capabilities of multimodal models. For example, the MME (Multimodal Multitask Evaluation) benchmark separately evaluates model performance scores in two dimensions—perception (such as image basic attribute recognition) and cognition (such as logical reasoning) [6]. MMBench adopts a CircularEval mechanism to reduce option bias in multiple-choice questions, achieving more stable accuracy evaluation [6]. In professional fields, medical multimodal evaluation benchmarks (such as MultiMedEval) provide unified testing tools covering medical image classification, lesion detection, and other tasks, and use medical diagnosis standard indicators such as precision, recall, and F1 score for measurement.

The core indicator system for deployment performance should go beyond model accuracy and construct a four-dimensional comprehensive evaluation framework:

Inference efficiency indicators: Including end-to-end latency (total processing time from input to output), system throughput (number of samples processed per unit time), and first-token latency (critical for interactive applications). In cloud-edge collaborative systems, end-to-end latency can be decomposed into the sum of latencies in multiple links such as encoding, transmission, inference, and decoding [25].

Resource utilization indicators: Covering memory footprint (total runtime memory), computational complexity (measured by FLOPs or Ops), and energy efficiency (number of samples processed per unit energy consumption). For example, the LUQ quantization method can reduce the memory footprint of the LLaVA-1.5 model by 40% compared to its 4-bit quantized version, while the NANOMIND framework achieves a 42.3% energy reduction on edge devices through hardware-software co-design [12].

Accuracy retention indicators: Mainly evaluate the accuracy loss introduced by optimization technologies, usually expressed as a percentage. In addition, different application scenarios have specific indicators; for example, medical diagnosis focuses on F1 score and AUC, while manufacturing inspection focuses on classification accuracy, false detection rate, and missed detection rate [15, 41].

Economic indicators: Evaluate the feasibility of deployment solutions from a commercial perspective, including cost savings percentage, Return on Investment (ROI), and marginal cost during system expansion. Studies have shown that cloud-edge collaborative deployment can save about 65% of costs through machine learning-assisted resource scheduling [27].

Table 10 summarizes this core indicator system, providing a basis for subsequent quantitative comparisons.

## 7.2. Quantitative Verification of Key Deployment Technologies

Different deployment technologies have varying impacts on the above indicators. Through the analysis of empirical studies from 2024 to 2026, the benefits and costs of various technologies can be clearly quantified.

Low-bit quantization technology significantly improves memory and computational efficiency but requires careful management of accuracy loss. The LUQ (Layerwise Ultra-Low Bit Quantization) method adopts a layered strategy, prioritizing ultra-low bit (e.g., 2-bit) quantization for layers with low activation entropy. On the LLaVA-1.5 model, this method achieves compression with an average bit width of 2.75 bits, reduces memory footprint by 40%, and controls accuracy loss within 10% on the MME benchmark [6]. The VecInfer method focuses on compressing the KV cache; its 2-bit quantization scheme achieves up to 8.3x end-to-end decoding acceleration on the Llama-3.1-8B model while maintaining performance comparable to full-precision models [24]. However, the choice of quantization format (INT vs. FP) is also crucial. Studies have shown that on modern architectures such as Blackwell, for 4-bit formats, FP (such as MXFP4) is generally superior to INT formats in algorithm accuracy; but when combined with outlier suppression technologies such as Hadamard transform, NVINT4 can outperform NVFP4 [7].

Heterogeneous MoE architecture and inference engine optimization jointly improve model computational efficiency and deployment flexibility. The expert parallelism (EP) and hybrid parallelism modes supported by TensorRT-LLM enable efficient deployment of large MoE models (such as Mistral Mixtral 8×7B) across multiple GPUs [8]. At the application level, heterogeneous MoE adapters designed for multimodal models can achieve performance comparable to full-parameter fine-tuning by only fine-tuning 5-8% of the parameters, greatly reducing the computational cost of deployment [42]. A unified resource-aware framework collaboratively optimizes mixed-precision quantization, sparse attention, and dynamic expert routing, achieving up to 2.4x end-to-end acceleration and 1.8x memory reduction on edge and embedded platforms while maintaining generation quality [43].

Cloud-edge-end collaborative deployment strategies achieve dynamic balance between latency, accuracy, and cost through dynamic scheduling and resource allocation. The DistMLLM framework adopts a multi-armed bandit online learning algorithm to dynamically allocate encoding and inference tasks to heterogeneous edge nodes, maximizing the long-term profits of all service providers while showing lower cumulative regret

**Table 11:** Performance Improvement Ranges of Key Deployment Technologies

| Deployment Technology | Inference Latency Improvement | Memory Footprint Reduction | Accuracy Loss | Cost Savings |
|---|---|---|---|---|
| 4-bit Quantization | 30 – 50% | 50 – 75% | 2 – 5% | 40 – 60% |
| 2-bit Quantization | 50 – 70% | 75 – 87.5% | 5 – 10% | 60 – 80% |
| Heterogeneous MoE Adaptation | 60 – 80% | 60 – 80% | 1 – 3% | 50 – 70% |
| TensorRT Optimization | 40 – 60% | 10 – 20% | 0 – 1% | 20 – 40% |
| Cloud-Edge Collaborative Scheduling | 30 – 50% | 20 – 40% | 1 – 3% | 30 – 50% |

than Round-Robin (RR) and Static Resource-Aware (SRA) baselines [25]. The AIVD framework achieves a balance between high-precision semantic generation and low latency in industrial visual inspection through collaboration between lightweight edge detectors and cloud-based large models. Its heterogeneous resource-aware dynamic scheduling algorithm maintains high throughput and stable latency under variable network conditions [26]. In medical edge scenarios, the intelligent Clinical Decision Support System (CDSS) based on multithreading achieves an inference speed improvement of up to 28% on devices such as Raspberry Pi and maintains a 100% prediction accuracy on specific datasets [14]. Table 11 comprehensively compares the typical performance improvement ranges brought by different deployment technologies, intuitively showing the trade-off space during technology selection.

### 7.3. Application-Oriented Evaluation Framework Design Principles

To effectively guide practical deployment, a unified and scalable evaluation framework needs to be established. The design of this framework should follow the following principles:

Standardized test environment: Evaluations must be conducted under clearly defined hardware and software configurations to ensure the comparability and reproducibility of results. This includes specifying edge devices (such as Jetson Orin series), server GPU models, operating system versions, drivers, and specific versions of inference engines (such as TensorRT, ONNX Runtime) [12, 25, 43].

Automated and modular process: The evaluation process should be automated and adopt a modular design to facilitate the integration of new models, hardware platforms, or optimization technologies. The process usually includes environment initialization, benchmark model testing, optimization technology application, multi-dimensional performance measurement (latency, memory, accuracy, energy consumption), and result analysis and report generation.

Scenario-driven indicator weighting: Different application scenarios have different priorities for performance indicators. In the evaluation framework, adjustable weights should be assigned to each indicator. For example, in real-time medical diagnosis, end-to-end latency and diagnosis accuracy (such as F1 score) should have the highest weights; while in offline manufacturing quality analysis, throughput and overall cost savings may be more critical [14, 15]. A comprehensive performance score (PS) can be expressed as a weighted sum of each standardized indicator and its weight, i.e., $PS = \sum_i w_i I_i$, where the weight $w_i$ is determined according to business requirements.

By constructing and adopting such a quantitative evaluation system, the industry can scientifically compare the advantages and disadvantages of different multimodal large model deployment solutions, providing a solid decision-making basis for technology selection and system optimization in specific scenarios, thereby promoting the landing and popularization of efficient, practical, and economical artificial intelligence solutions in key fields such as medical care and manufacturing.

## 8. Future Development Trends and Technical Challenges Outlook

The deployment and application of multimodal large models in core fields such as healthcare and smart manufacturing are in a critical stage of transitioning from technical breakthroughs to large-scale implementation. Looking back at the development process from 2024 to 2026, we have witnessed the rapid evolution of deployment technologies from theoretical verification to engineering practice. However, the path to efficient, inclusive, and reliable deployment is still full of systematic challenges. Based on the current technical status, this section will analyze the future evolution direction of multimodal large model deployment, evaluate the impact of emerging paradigms such as neuromorphic computing and quantum computing, and deeply explore deep-seated obstacles such as standardization, interoperability, and ecosystem construction.

### 8.1. Neuromorphic Computing: Reshaping the Energy Efficiency Paradigm of Edge Deployment

Neuromorphic computing, which simulates the parallel, sparse, and event-driven information processing mechanism of the human brain, provides a revolutionary idea for solving the high energy consumption problem of multimodal model deployment on edge devices. In 2025, Intel's released Hala Point system marked the transition of this technology from the laboratory to practical application. The system integrates 1.15 billion neurons and 1152 Loihi 2 processors, achieving 20 quadrillion operations per second in a volume equivalent to a microwave oven, and achieving significant results of 50x acceleration and 100x energy savings on specific optimization problems [44]. This energy efficiency advantage is crucial for medical edge devices that need to work continuously for a long time (such as wearable health monitors), enabling up to 20.8 hours of event-triggered inference [12].

The core value of neuromorphic architecture in multimodal scenarios lies in its adaptive processing capability for heterogeneous data streams. In medical diagnosis, medical imaging data is dense, while physiological signals (such as ECG and EEG) have high sparsity and temporal correlation. Traditional computing architectures have a lot of redundant computations when processing such mixed data streams. Through simulating synaptic plasticity and neuron spiking mechanisms, neuromorphic chips can dynamically allocate computing resources according to the characteristics of input data, realizing "on-demand activation", thereby significantly reducing average power consumption. The BI Explorer-1

(BIE-1) neuromorphic server developed by the Guangdong Institute of Intelligent Science and Technology further verifies this potential. While maintaining a 90% power reduction, it achieves a training speed of 100,000 tokens per second and an inference speed of 500,000 tokens per second [45].

However, the practical deployment of neuromorphic computing still faces severe challenges. First, its software ecosystem is not yet mature, and the integration of programming models with mainstream deep learning frameworks (such as PyTorch and TensorFlow) is limited, requiring developers to master new programming paradigms. Second, the collaborative scheduling algorithm with existing heterogeneous computing units such as CPU, GPU, and NPU is complex, making it difficult to achieve seamless cross-architecture computing migration in multimodal task flows. Finally, the manufacturing cost of neuromorphic chips is high, and there is a lack of unified design and interface standards, limiting their popularization in cost-sensitive edge scenarios.

### 8.2. Quantum-Classical Hybrid Computing: A New Path for Solving Complex Optimization Problems

The integration of quantum computing and classical artificial intelligence is moving from the ”advantage demonstration” stage to the practical exploration period of ”Noisy Intermediate-Scale Quantum (NISQ)” [44]. For multimodal model deployment, quantum computing is not intended to completely replace classical computing, but to provide acceleration for specific high-complexity problems in model training and inference optimization through quantum-classical hybrid architectures.

Its application potential is mainly reflected in two aspects: first, accelerating the solution of complex optimization problems, such as finding the globally optimal task offloading and computing allocation scheme in cloud-edge collaboration; second, in the multimodal feature fusion stage, quantum machine learning algorithms may more efficiently process high-dimensional and nonlinear inter-modal correlations [44]. The industry has emerged systems such as DGX Quantum jointly launched by Quantum Machines and NVIDIA, which aim to closely couple quantum controllers with classical AI superchips through microsecond-level latency, realizing real-time quantum error correction and improving the stability and practicality of quantum operations [44]. Nevertheless, the application of quantum computing in multimodal deployment is still in its early stages. Major obstacles include insufficient stability and scalability of quantum hardware, high technical barriers to integrating quantum algorithms with existing deep learning workflows, and the acquisition and operation costs of quantum computing resources far exceeding the budgets of most enterprises. Industry forecasts indicate that hybrid quantum-AI services for specific fields (such as drug discovery and material simulation) may achieve initial commercialization between 2026 and 2027, but their widespread application in general multimodal inference tasks still requires time [44].

### 8.3. Standardization and Interoperability: The Cornerstone of Large-Scale Deployment

Currently, there are serious fragmentation problems in all links of multimodal model deployment, which have become the core bottleneck hindering large-scale technology promotion. This is mainly reflected in multiple levels such as model format, quantization parameters, inference engine interfaces, and cross-modal data representation.

First, model formats and quantization standards are not unified. Different manufacturers and open-source communities (such as TensorRT and ONNX Runtime) adopt different quantization schemes, parameter storage formats, and calibration methods. This leads to the need for repeated quantization tuning when deploying the same model on different hardware platforms (such as cloud A100 and edge Jetson devices), which is not only inefficient but also may introduce additional accuracy loss [43]. Unifying quantization standards is the key to reducing deployment complexity.

Second, the interoperability between inference engines and runtime environments is insufficient. Although there are multiple high-performance inference engines, their APIs, plug-in mechanisms, and memory management strategies lack unified abstraction. For example, in the collaborative inference framework AIVD, the collaboration between large and small models relies on a specific runtime environment, limiting its flexible migration between heterogeneous platforms [26]. This increases the complexity and maintenance cost of system integration.

Third, cross-modal data representation and exchange standards are missing. In professional fields such as medical care, different modal data (such as DICOM images, HL7 text reports, and IEEE 11073 physiological signals) follow independent industry standards. The lack of unified cross-modal data representation and preprocessing pipeline standards makes each deployment project require a customized data conversion layer, hindering the construction and reuse of standardized solutions.

Finally, cloud-edge-end collaboration protocols lack interoperability. Edge computing solutions from major cloud service providers (AWS, Azure, etc.) mostly adopt proprietary task offloading and resource management protocols, forming technical barriers. This makes it extremely complex to build cross-vendor and cross-platform hierarchical collaborative deployment architectures (such as remote medical diagnosis systems), and users are easily locked into a single ecosystem [27].

### 8.4. Ecosystem Construction: The Bridge from Technology to Industry

A healthy and prosperous ecosystem is a necessary condition for multimodal AI deployment technology to move from the laboratory to industrial applications. Current ecosystem construction faces multiple challenges such as developer toolchains, industry solutions, and talent training.

Developer toolchains urgently need integration and improvement. The deployment process of multimodal models involves multiple steps such as model conversion, quantization, compilation, and performance analysis. Currently, there is a lack of an end-to-end unified tool platform. Developers need to combine independent tools from different manufacturers, facing a steep learning curve and high tool switching costs. For example, TensorRT-Edge-LLM is mainly optimized for language models and has limited support for complex multimodal scenarios [46].

Industry solutions lack standardized reference architectures. Vertical fields such as medical care and manufacturing have clear business scenarios and technical requirements (such as data privacy compliance, real-time requirements, and system integration), but each project often designs the deployment architecture from scratch, resulting in a lot of redundant work and resource waste. Establishing an industry knowledge base containing best practices, reference implementations, and compliance inspection tools will greatly accelerate the landing process of applications.

Interdisciplinary talents are in severe shortage. Multimodal model deployment requires the integration of professional knowledge in deep learning, edge computing, signal processing, and specific fields (such as medicine and industrial engineering). The existing education system often

separates these disciplines, leading to talents with single knowledge structures who are difficult to cope with complex engineering challenges. Cultivating interdisciplinary talents who understand both AI algorithms and domain business and engineering deployment is the core of the sustainable development of the ecosystem.

### 8.5. Systematic Challenges: Multidimensional Considerations Beyond Technology

The obstacles to multimodal large model deployment are not limited to technical aspects but also interweave systematic challenges such as economic feasibility and ethical compliance, which need to be comprehensively balanced in technology route selection.

Economic feasibility and Return on Investment (ROI) are the core of decision-making. Deployment costs include not only hardware procurement but also software licenses, system integration, long-term operation and maintenance, and personnel training. Although the global AI medical edge computing market is expected to grow from 2.292 billion US dollars in 2024 to 4.619 billion US dollars by 2032 [47], the budget of individual medical institutions or factories is limited. It is necessary to accurately evaluate the efficiency improvement, cost savings, and revenue growth brought by deployment to determine a reasonable ROI cycle.

Data privacy and security compliance are the red lines of deployment. Medical data is strictly protected by regulations such as HIPAA and GDPR, and industrial data involves trade secrets. Deployment solutions must ensure localized processing of sensitive data, encrypted transmission, and access control. Currently, the protection mechanisms of privacy computing technologies such as federated learning in the inference stage are still immature, and the performance and energy consumption of strong privacy technologies such as homomorphic encryption on edge devices are still bottlenecks.

Model interpretability and decision credibility are key requirements. In high-risk scenarios such as medical diagnosis, physicians and engineers not only need model outputs but also need to understand the reasoning basis. The decision-making of multimodal models integrates multiple information sources, and their internal logic is more complex than single-modal models. Existing interpretable AI technologies (such as attention mechanism visualization) have limited effects in multimodal fusion scenarios, and there is an urgent need to develop specialized interpretability methods to establish user trust.

Technical debt and long-term maintenance risks are often underestimated. Multimodal deployment systems involve complex software and hardware dependency stacks. With rapid technological iteration and changes in business requirements, system maintenance costs may rise sharply. Especially in the field of medical devices that require long-term certification, the contradiction between the technical update cycle and the certification cycle is particularly prominent. It is necessary to find a balance between technical advancement and long-term system stability.

### 8.6. Technology Evolution Roadmap (2026-2030 Outlook)

Based on current trends and challenges, the evolution path of multimodal model deployment technology in the next few years can be outlined:

1. 2026-2027: Standardization breakthroughs and hybrid architecture maturity period. It is expected that intermediate representation formats such as ONNX will extend unified support for multimodal models and quantization parameters. Quantum-classical hybrid computing will achieve initial commercial application in specific optimization tasks such as material simulation and drug discovery. Neuromorphic chips will begin to be integrated into low-power wearable devices, achieving significant improvements in energy efficiency [44].
2. 2028-2029: Adaptive deployment and ecosystem prosperity period. Dynamic resource scheduling systems based on reinforcement learning will become mainstream, capable of adaptively adjusting computing strategies according to real-time loads and hardware status. Cross-modal federated learning frameworks will mature, supporting collaborative optimization among multiple institutions while protecting privacy. Open-source communities and commercial solutions around multimodal deployment will form positive interactions, spawning a number of innovative start-ups.
3. 2030 and beyond: Inclusive AI and deepened sustainable development period. The threshold for deployment technology will be significantly reduced, enabling small and medium-sized institutions to afford advanced AI solutions. Green computing, energy efficiency optimization, and electronic waste management will be deeply integrated into the technology life cycle. Quantum computing is expected to show unique advantages in more multimodal tasks that need to process ultra-high-dimensional and strongly correlated data, and cutting-edge interaction technologies such as brain-computer interfaces may open up new human-machine collaboration models.

Realizing this evolution path requires close cooperation between academia, industry, and policymakers. Only through continuous cross-domain innovation, open cooperation, and systematic planning can we overcome current obstacles and ultimately release the full potential of multimodal large models in promoting social progress and industrial upgrading.

## 9. Conclusion

This survey systematically reviews the technical evolution, core innovations, and industrial application practices of multimodal large model (MLLM) efficient deployment from 2024 to 2026, focusing on three core pillars: heterogeneous Mixture-of-Experts (MoE) architecture, low-bit quantization, and cloud-edge-end collaborative deployment. By integrating the latest research results and practical cases, this work constructs a comprehensive technical framework for MLLM deployment, providing valuable references for bridging the gap between theoretical model capabilities and industrial implementation needs.

Heterogeneous MoE architecture has emerged as a key paradigm for balancing multimodal adaptation and computational efficiency. Through modality-isolated routing, router orthogonal loss, and specialized expert division, it addresses the limitations of traditional homogeneous architectures in multimodal feature processing, achieving refined resource allocation while maintaining high model performance. Low-bit quantization technologies represented by Layerwise Ultra-Low Bit Quantization (LUQ) and MoEQuant have broken through the bottleneck of multimodal data quantization, reducing model memory footprint by 31.5%-40% while controlling accuracy loss within 10%, laying the foundation for MLLM deployment on resource-constrained edge devices. Cloud-edge-end collaborative strategies such as DistMLLM, AIVD, and NANOMIND have realized intelligent task offloading and dynamic resource scheduling, meeting the dual requirements of low latency and

high privacy in medical diagnosis, smart manufacturing, and other fields, while achieving significant energy savings (up to 42.3%) and cost reductions (25%-67%).

The application cases in healthcare and smart manufacturing fully verify the practical value of these deployment technologies. In medical edge diagnosis, lightweight MLLMs deployed on Raspberry Pi and other devices achieve 100% prediction accuracy and 28% inference speed improvement, complying with HIPAA and FDA regulatory requirements. In smart manufacturing, predictive maintenance and real-time quality inspection systems based on multimodal data fusion increase defect identification accuracy to 94% and reduce equipment downtime by over 30%, demonstrating remarkable economic benefits with a three-year ROI of 184%. These achievements confirm that MLLM efficient deployment technologies have effectively solved industry pain points such as high latency, high resource consumption, and data privacy risks.

Despite significant progress, MLLM deployment still faces systemic challenges. The fragmentation of model formats, quantization standards, and inference engine interfaces hinders cross-platform interoperability; model interpretability and decision transparency remain key barriers to adoption in high-risk scenarios such as medical care; and the balance between privacy protection technologies (e.g., federated learning, homomorphic encryption) and computational efficiency needs further optimization. Additionally, the shortage of interdisciplinary talents and the complexity of technical integration restrict the large-scale promotion of deployment solutions.

Looking forward, the development of MLLM efficient deployment will focus on adaptive, standardized, and sustainable directions. Adaptive deployment architectures based on reinforcement learning will dynamically adjust model configurations according to real-time resource status and task requirements; the unification of cross-modal data representation standards and inference engine interfaces will reduce deployment complexity; the deep integration of neuromorphic computing, quantum-classical hybrid computing, and MLLMs will reshape the energy efficiency and computational capabilities of edge deployment. Furthermore, the combination of privacy-enhancing computing and trusted execution environments will provide stronger security guarantees for sensitive fields, while the construction of a complete ecosystem including toolchains, reference architectures, and talent training systems will accelerate the industrialization of MLLM deployment technologies.

In summary, the efficient deployment of multimodal large models is a core link in realizing the industrial value of artificial intelligence. Through the continuous innovation and integration of heterogeneous MoE, low-bit quantization, and cloud-edge-end collaboration technologies, and the joint efforts of academia, industry, and policymakers to address standardization, interpretability, and ecosystem challenges, MLLMs will be more widely and reliably applied in edge terminals, industrial sites, and medical institutions, becoming a key driving force for promoting intelligent upgrading in various industries and achieving inclusive intelligence.

## References


[1] Liang, C. X., Tian, P., Yin, C. H., et al. (2024). *A comprehensive survey and guide to multimodal large language models in vision-language tasks*. arXiv preprint. https://doi.org/10.48550/arXiv.2411.06284

[2] Technode feed. (2025, June 30). *Baidu open-sources ERNIE 4.5 multimodal AI model*. technode. https://00l.xyz/03U3s

[3] Baidu Inc. (2025). *ERNIE 4.5: Multimodal large language model*. https://github.com/PaddlePaddle/ERNIE

[4] Koparkar, S. (2025, December 3). *Mixture of experts powers the most intelligent frontier AI models, runs 10x faster to deliver 1/10 the token cost on NVIDIA Blackwell NVL72*. NVIDIA Blog. https://blogs.nvidia.com/blog/mixture-of-experts-blackwell-nvl72

[5] Dutt, R., Hanspal, H., Xia, G., et al. (2025). *Exploiting mixture-of-experts redundancy unlocks multimodal generative abilities*. arXiv preprint. https://doi.org/10.48550/arXiv.2503.22517

[6] Bhatnagar, S., Xu, A., Tan, K.-H., et al. (2025). *LUQ: Layerwise ultra-low bit quantization for multimodal large language models*. arXiv preprint. https://arxiv.org/abs/2509.23729

[7] Chen, M., Wu, M., Jin, H., et al. (2025). *INT v.s. FP: A comprehensive study of fine-grained low-bit quantization formats*. arXiv preprint. https://arxiv.org/abs/2510.25602

[8] Fedus, W., Zoph, B., Shazeer, N., et al. (2022). Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research, 23*(120), 1–39. http://jmlr.org/papers/v23/21-0998.html

[9] ONNX Runtime. (2025). *TensorRT execution provider*. ONNX Runtime Documentation. https://onnxruntime.ai/docs/execution-providers/TensorRT-ExecutionProvider.html

[10] Mao, H. Z., Xin, J., Yu, Y., et al. (2025). *Enable Blackwell inference with TensorRT model optimizer*. NVIDIA On-Demand. https://www.nvidia.com/en-us/on-demand/session/gtc25-s72609/

[11] Qiu, H., Biswas, A., Zhao, Z., et al. (2025). *MoDServe: Modality- and stage-aware resource disaggregation for scalable multimodal model serving*. arXiv preprint. https://arxiv.org/abs/2502.00937

[12] Li, Y., Zhang, S., Zeng, Y., et al. (2025). *Tiny but mighty: A software-hardware co-design approach for efficient multimodal inference on battery-powered small devices*. arXiv preprint. https://arxiv.org/abs/2510.05109

[13] Longpre, S., Akiki, C., Lund, C., et al. (2025). *Economies of open intelligence: Tracing power & participation in the model ecosystem*. arXiv preprint. https://doi.org/10.48550/arXiv.2512.03073

[14] Alghareeb, F. S., & Hasan, B. T. (2025). Leveraging multithreading on edge computing for smart healthcare based on intelligent multimodal classification approach. *Computerized Medical Imaging and Graphics, 124*, 102594. https://doi.org/10.1016/j.compmedimag.2025.102594

[15] Rydziński, S., Zahradnikova, B., Sutova, Z., et al. (2024). A predictive quality inspection framework for the manufacturing process in the context of Industry 4.0. *Sensors, 24*(17), 5644. https://doi.org/10.3390/s24175644

[16] ERNIE Bot Team (Baidu Inc.). (2025, June 30). *Announcement of open source release of the ERNIE 4.5 model family*. Baidu Research Blog. https://ernie.baidu.com/blog/posts/ernie4.5/

[17] ERNIE Team, Baidu Inc. (2025). *ERNIE 4.5 technical report*. https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

[18] Hu, X., Chen, Z., Yang, D., et al. (2025). *MoEQuant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance*. arXiv preprint. https://doi.org/10.48550/arXiv.2505.03804

[19] Fang, Z., Lin, Z., Hu, S., et al. (2025). *HFedMoE: Resource-aware heterogeneous federated learning with mixture-of-experts*. arXiv preprint. https://arxiv.org/abs/2601.00583

[20] Bhatnagar, S., Xu, A., Tan, K.-H., et al. (2025). *LUQ: Layerwise ultra-low bit quantization for multimodal large language models*. arXiv preprint. https://doi.org/10.48550/arXiv.2509.23729

[21] NVIDIA. (2025). *TensorRT LLM: NVIDIA's open-source library for accelerating large language model inference*. NVIDIA Developer Documentation. https://developer.nvidia.com/tensorrt-llm

[22] NVIDIA. (2025). *NVIDIA/TensorRT-LLM - GitHub*. GitHub Repository. https://github.com/NVIDIA/TensorRT-LLM

[23] ONNX Runtime. (2025). *Quantize ONNX models*. ONNX Runtime Documentation. https://onnxruntime.ai/docs/performance/model-optimizations/quantization.html

[24] Yao, D., Yang, C., Tong, Z., et al. (2025). *VecInfer: Efficient LLM inference with low-bit KV cache via outlier-suppressed vector quantization*. arXiv preprint. https://arxiv.org/abs/2510.06175

[25] Yuan, X., Chen, H., Liu, L., et al. (2025). DistMLLM: Enhancing multimodal large language model serving in heterogeneous edge computing. *Sensors, 25*(24), 7612. https://doi.org/10.3390/s25247612

[26] Hu, Y., Yang, Z., Zhao, C., et al. (2025). *AIVD: Adaptive edge-cloud collaboration for accurate and efficient industrial visual detection*. arXiv preprint. https://arxiv.org/abs/2601.04734v1

[27] Pasdar, A., Lee, Y. C., Hassanzadeh, T., et al. (2021). Resource recommender for cloud-edge engineering. *Information, 12*(6), 224. https://doi.org/10.3390/info12060224
[28] Wang, Z., Zhao, H., Yang, Y., et al. (2025). *OMNIFUSER: Adaptive multimodal fusion for service-oriented predictive maintenance*. arXiv preprint. https://arxiv.org/abs/2511.01320
[29] Kang, S. (2025). *Trustworthy equipment monitoring via cascaded anomaly detection and thermal localization*. arXiv preprint. https://arxiv.org/abs/2512.24755
[30] Lueh, K. L. (2025, September 9). *Industrial AI market: 10 insights on how AI is transforming manufacturing*. Market Research Report. https://iot-analytics.com/industrial-ai-market-insights-how-ai-is-transforming-manufacturing/
[31] Li, X. (2025, September 1). *Artificial intelligence empowers the digital-intelligent transformation of China's manufacturing industry*. People's Tribune, 04. https://paper.people.com.cn/rmlt/pc/content/202509/01/content_30104680.html
[32] Putteti, S., Santhi, G., Mittoor, G. R., et al. (2025). Intelligent industrial IoT: A data-driven approach for smart manufacturing and predictive maintenance. In *2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)* (pp. 1032–1040). https://doi.org/10.1109/ICAISS61471.2025.11041978
[33] Prasser, D. R. (2025, July 21). *Future of manufacturing: 13 trends driving 2026-2035 growth*. StartUs Insights Blog. https://www.startus-insights.com/innovators-guide/future-of-manufacturing/
[34] Zhang, W., Yang, D., Gao, S., et al. (2026). Conclusions. In *Intelligent resource scheduling in end-edge-cloud networks* (pp. 131–152). Springer, Cham. https://doi.org/10.1007/978-3-032-07667-0_6
[35] Yu, W., Liu, Y., Dillon, T., et al. (2023). Edge computing-assisted IoT framework with an autoencoder for fault detection in manufacturing predictive maintenance. *IEEE Transactions on Industrial Informatics, 19*(4), 5701–5710. https://doi.org/10.1109/TII.2022.3178732
[36] Fang, Z., Lin, Z., Hu, S., et al. (2026). *HFedMoE: Resource-aware heterogeneous federated learning with mixture-of-experts*. arXiv preprint. https://doi.org/10.48550/arXiv.2601.00583
[37] Wang, X., Li, X., Wang, N., et al. (2022). Fine-grained cloud edge collaborative dynamic task scheduling based on DNN layer-partitioning. In *2022 18th International Conference on Mobility, Sensing and Networking (MSN)* (pp. 155–162). https://doi.org/10.1109/MSN57253.2022.00037
[38] Doslouglu, T., & MacDonald, M. (2022). *Circuit design for predictive maintenance*. arXiv preprint. https://doi.org/10.48550/arXiv.2211.10248
[39] Zhao, F., Wu, Y., Hu, M., et al. (2025). Current progress of digital twin construction using medical imaging. *Journal of Applied Clinical Medical Physics, 26*(8), e70226. https://doi.org/10.1002/acm2.70226
[40] Wang, K., Zhou, X., & Guan, J. (2026). The construction of an integrated cloud network digital intelligence platform for rail transit based on artificial intelligence. *Scientific Reports, 16*, 393. https://doi.org/10.1038/s41598-026-39968-2
[41] Hao, Y., Cheng, C., Li, J., et al. (2025). Multimodal integration in health care: Development with applications in disease management. *Journal of Medical Internet Research, 27*, e76557. https://doi.org/10.2196/76557
[42] Zhou, S., Xia, Y., Huang, H., et al. (2025). Enhancing multi-modal models with heterogeneous MoE adapters for fine-tuning. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). https://doi.org/10.1109/ICME59968.2025.11210043
[43] Wang, Y., & Zhao, J. (2025). A unified and resource-aware framework for adaptive inference acceleration on edge and embedded platforms. *Electronics, 14*(11), 2188. https://doi.org/10.3390/electronics14112188
[44] Shimabukuro, J. (2025, October 28). *Under-radar AI disruptors (projections from late-Oct. 2025)*. ETC Journal. https://etcjournal.com/2025/10/28/under-radar-ai-disruptors-projections-from-late-oct-2025/
[45] Laurent, A. (2025, December). *Latest AI research (Dec 2025): GPT-5, agents & trends*. Intuition Labs Blog. https://intuitionlabs.ai/articles/latest-ai-research-trends-2025
[46] NVIDIA. (2025). *Overview - TensorRT Edge-LLM*. NVIDIA Documentation. https://nvidia.github.io/TensorRT-Edge-LLM/0.4.0/developer_guide/01.1_Overview.html
[47] IntelMarket Research. (2025, September 16). *AI medical edge computing system market growth analysis, dynamics, key players and innovations, outlook and forecast 2025-2032*. Market Research Report. https://www.intelmarketresearch.com/ai-medical-edge-computing-system-market-7571