Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

Cross-Terminal Intelligent Diagnosis and Treatment System Based on Multimodal Large Language Models

Li Yuyuan*, Shi Jingang, Li Xiaolei, Liu Xinyu, Lang Shouhe

Shenyang Jianzhu University, School of Computer Science and Engineering, Hunnan District, Shenyang, Liaoning 110170, China

*Corresponding to: Li Yuyuan

Abstract: To address the prominent challenges in current medical auxiliary diagnosis—including inaccurate tongue image feature extraction, poor disease adaptability, lack of cross-terminal collaboration, and high dependence on foreign core technologies—a cross-terminal intelligent diagnosis and treatment system based on multimodal large language models was designed and implemented. The system takes traditional Chinese medicine (TCM) tongue diagnosis as the core application scenario. Using SAM-2 with LoRA lightweight fine-tuning, pixel-level precise segmentation of tongue images is achieved with 97.2% accuracy. A heterogeneous fusion feature extraction architecture combining ResNet and Vision Transformer is proposed, enabling three-layer information fusion of tongue body, tongue coating, and tongue texture, improving disease prediction accuracy to 84%. The Qwen3-VL multimodal large language model integrated with Retrieval-Augmented Generation (RAG) technology constructs an interpretable disease prediction engine with a retrieval precision rate of 61%. Full-stack deployment is completed on the domestic Kunpeng CPU and Ascend NPU hardware platform, achieving an inference speed of 20 Token/s. Experimental results demonstrate that the system achieves significant performance in accuracy, interpretability, and domestic adaptation, validating the feasibility and efficiency of domestic hardware and software systems in handling complex multimodal large model tasks.

Keywords: general medical auxiliary diagnosis; image segmentation; multimodal large language models; retrieval-augmented generation; domestic adaptation; cross-terminal collaboration

1 INTRODUCTION

Driven by the "Artificial Intelligence+" strategy and the wave of digital health, AI technology has deeply penetrated the medical field. Machine learning and deep learning have brought breakthroughs to medical assisted diagnosis, promoting the transformation toward intelligent and standardized diagnosis while alleviating pain points such as uneven medical resource distribution and low efficiency. With the emergence of algorithms such as SAM-2[1], the Qwen3 series[2], ResNet[3], and Vision Transformer[4], related technologies have demonstrated excellent performance in pneumonia detection and TCM tongue diagnosis. However, current intelligent medical assisted diagnosis still faces numerous shortcomings: insufficient accuracy in medical image feature extraction, difficulty in multimodal data adaptation, lack of cross-terminal collaboration, high dependence on foreign core technologies and hardware, inadequate domestic substitution, and difficulty in ensuring data security. Current medical imaging data grows at over 30% annually. Traditional manual film reading is inefficient (15-30 minutes per CT case), with a misdiagnosis rate of approximately 15%, and physician diagnostic variability as high as 20%, making it difficult to meet massive data demands[5]. The medical industry's requirements for comprehensiveness, accuracy, and security of systems are increasingly demanding. Researching and developing a general medical aided diagnosis system with comprehensiveness, accuracy, domestic substitution, and cross-terminal collaboration capability has significant theoretical and practical value.

In recent years, algorithms such as SAM-2, the Qwen3 series, ResNet, and Vision Transformer have emerged and matured, demonstrating excellent performance in pneumonia detection and TCM tongue diagnosis. However, clinical applications still face bottlenecks including inaccurate lesion extraction, insufficient adaptability of various technologies, and algorithm "black box" issues, with physician decision-making trust as low as 43%, severely constraining large-scale deployment. Fine-tuning techniques such as LoRA[6] have become key to deploying large models in vertical domains, while RAG technology[7] has become the core pathway for knowledge

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

integration in vertical domains. Domestic scholars have leveraged models such as SAM-2, ResNet, and Transformers to form differentiated advantages in TCM tongue diagnosis[8] and rare disease diagnosis, proposing a fusion architecture of ResNet and Transformers to optimize medical image segmentation and feature extraction algorithms.

In cross-terminal collaboration, different technical pathways have emerged both domestically and internationally around three core objectives: device interconnection, data synchronization, and remote collaboration. International approaches center on "cloud collaboration + multi-device adaptation," building a full-scenario collaboration system covering wearable devices, edge terminals, and cloud platforms. Domestic research is grounded in tiered diagnosis needs, leveraging desktop cloud technology to build integrated digital workspaces, enabling seamless roaming and data synchronization for physicians across different terminals. In medical domestic adaptation, domestic efforts focus on Kunpeng CPUs, Ascend NPUs, and other domestic hardware to build a domestic medical server foundation. Huawei's DaVinci architecture adopts 3D Cube matrix multiplication units, capable of completing 4096 MAC operations in a single clock cycle, significantly optimizing the efficiency of medical large model training and pathological feature extraction.

This paper addresses the aforementioned technical bottlenecks by designing and implementing a cross-terminal intelligent diagnosis and treatment system based on multimodal large language models. The system innovates around four core algorithm modules—image segmentation, feature extraction, disease prediction, and retrieval—and completes full-stack domestic deployment and engineering validation based on Kunpeng CPUs and Ascend NPUs.

2 System Architecture Design

The core application scenario of this system focuses on the field of TCM tongue diagnosis. The overall architecture is built around four core objectives: "accuracy, efficiency, interpretability, and clinical adaptability," while also accommodating general medical image processing needs. The core algorithm of the system is divided into four stages, each working in coordination and progressing step by step:

(1) Image Segmentation Stage: The SAM-2 algorithm combined with LoRA lightweight fine-tuning is employed to achieve pixel-level precise segmentation of tongue images, separating the tongue coating and background regions, capturing detailed tongue features, and providing high-purity input sources for subsequent feature extraction.

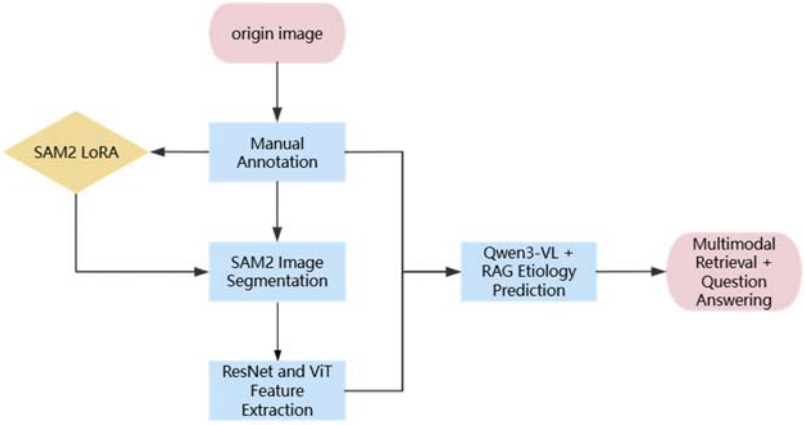

Fig. 1 System Architecture Diagram

(2) Image Feature Extraction Stage: A heterogeneous fusion network of ResNet and Vision Transformer is used to extract features from original images, tongue coating images, and tongue texture images in a mixed manner. Through feature concatenation and multi-scale feature pooling, multi-dimensional features are mapped into a unified TCM diagnostic feature space.

(3) Disease Prediction Stage: The Qwen3-VL multimodal large language model is integrated with Retrieval-Augmented Generation (RAG) technology, combining tongue features with clinical text information to achieve accurate prediction of TCM syndrome types and disease causes. A knowledge base self-verification mechanism enhances diagnostic interpretability.

(4) Disease Retrieval Stage: Based on VL-Embedding vector embedding technology and RAG retrieval-augmented technology, a case vector database is constructed to quickly retrieve similar cases and treatment plans, assisting physicians in optimizing diagnostic decisions.

The system simultaneously builds a domestic hardware foundation based on "Kunpeng CPU + Ascend NPU," centered on the "hardware-software-algorithm" trinity adaptation philosophy, achieving full-stack domestic deployment.

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

3 KEY ALGORITHM DESIGN

3.1 IMAGE SEGMENTATION MODULE

The image segmentation module serves as the core preprocessing stage of this system, aiming to achieve pixel-level precise segmentation of TCM tongue images. By removing background noise such as oral mucosa, teeth, and lips, high-fidelity tongue target regions are extracted. To address the limitations of traditional segmentation algorithms in edge blurring and detail loss, this system adopts SAM-2 (Segment Anything Model 2) as the pre-trained backbone network and innovatively introduces LoRA (Low-Rank Adaptation) parameter-efficient fine-tuning.

The core architecture of SAM-2 consists of three modules: an image encoder, a prompt encoder, and a mask decoder. The image encoder adopts an improved Vision Transformer (ViT) architecture, capturing both overall morphology and local detail features of the tongue through multi-scale feature extraction. The prompt encoder processes user-input prompt information, guiding the model to focus on target feature regions. The mask decoder generates precise masks based on the fused features. To achieve effective multi-scale feature fusion, the following formula is adopted:

$$F_{fusion} = \sum_{k=1}^{K} \omega_k \cdot F_k \quad (1)$$

Where F represents the fused final feature map, K is the number of feature scales, w_k represents the weight coefficient for the k-th scale feature (satisfying Σw_k=1), and F_k represents the image feature map at the k-th scale. This formula achieves efficient fusion of features at different resolutions through adaptive weight allocation.$F_{\text{fusion}}\omega_k \sum_{k=1}^{K} \omega_k = 1 F_k$

FIG. 2 COMPARISON OF ORIGINAL, ANNOTATED, AND PREDICTED IMAGES

This module embeds LoRA lightweight adapters into the self-attention layers of Hiera-ViT. The core theory of LoRA is based on the "low intrinsic rank hypothesis," which suggests that although pre-trained models have over-parameterized characteristics, their weight updates have a low intrinsic rank when adapting to downstream tasks. Based on this, the pre-trained weights are frozen, and two trainable low-rank decomposition matrices A and B are injected to approximate the weight update process. Subsequent improved variants such as QLoRA[10] have further enhanced fine-tuning efficiency by incorporating quantization techniques:

$$h = W_0 x + \Delta W x = W_0 x + \frac{\alpha}{r} BAx \quad (2)$$

Where x is the input vector, W₀ is the frozen pre-trained weight matrix, α is the scaling coefficient, and r is the rank of the low-rank matrices. This mechanism requires training only approximately 10% of the parameters to enable the model to accurately capture micro-semantic features such as tongue edge details, tongue texture color, and tongue coating texture. During initialization, matrix A follows a random Gaussian distribution, and matrix B is initialized as a zero matrix, ensuring that the model behavior at the start of training is completely consistent with the original pre-trained model. Experiments show that the fine-tuned model achieves a segmentation accuracy (PA) of 97.2%, an improvement of 4.8 percentage points over the un-fine-tuned SAM-2, with a single image segmentation time of 600ms.

3.2 IMAGE FEATURE EXTRACTION MODULE

The feature extraction module receives the output of the image segmentation module. Its core function is to extract features relevant to TCM diagnosis from the precisely segmented tongue images, including tongue texture color, tongue coating thickness, tongue body morphology, and sublingual vessels. This module adopts a heterogeneous fusion architecture of CNN (ResNet) and Transformer (ViT), leveraging the complementarity of the two mainstream visual backbone networks to achieve optimal feature extraction.

In terms of local detail capture, ResNet leverages the inductive bias characteristics of convolutional neural networks, effectively solving the vanishing gradient problem through deep residual connections, precisely capturing fine-grained local pathological features such as tongue coating texture, papules, and petechiae. Its residual block formula is:

$$y = F(x, W) + x \quad (3)$$

Where x is the input feature of the residual block, and F(x, {W_i}) is the mapping function of the residual block's main path.

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

When the network is deep, if the main path mapping approaches 0, the output y is approximately equal to x, achieving identity mapping and effectively alleviating the vanishing gradient problem.F(x,W)F(x,W)

In terms of global context modeling, ViT utilizes the self-attention mechanism to break through local receptive field limitations, modeling tongue morphology (e.g., enlarged, thin) and the overall spatial distribution of tongue texture and coating from a global perspective. The core process of ViT includes four steps: image patchification, embedding encoding, positional encoding, and Transformer encoding. The multi-head attention mechanism captures the relationships between different image patches.

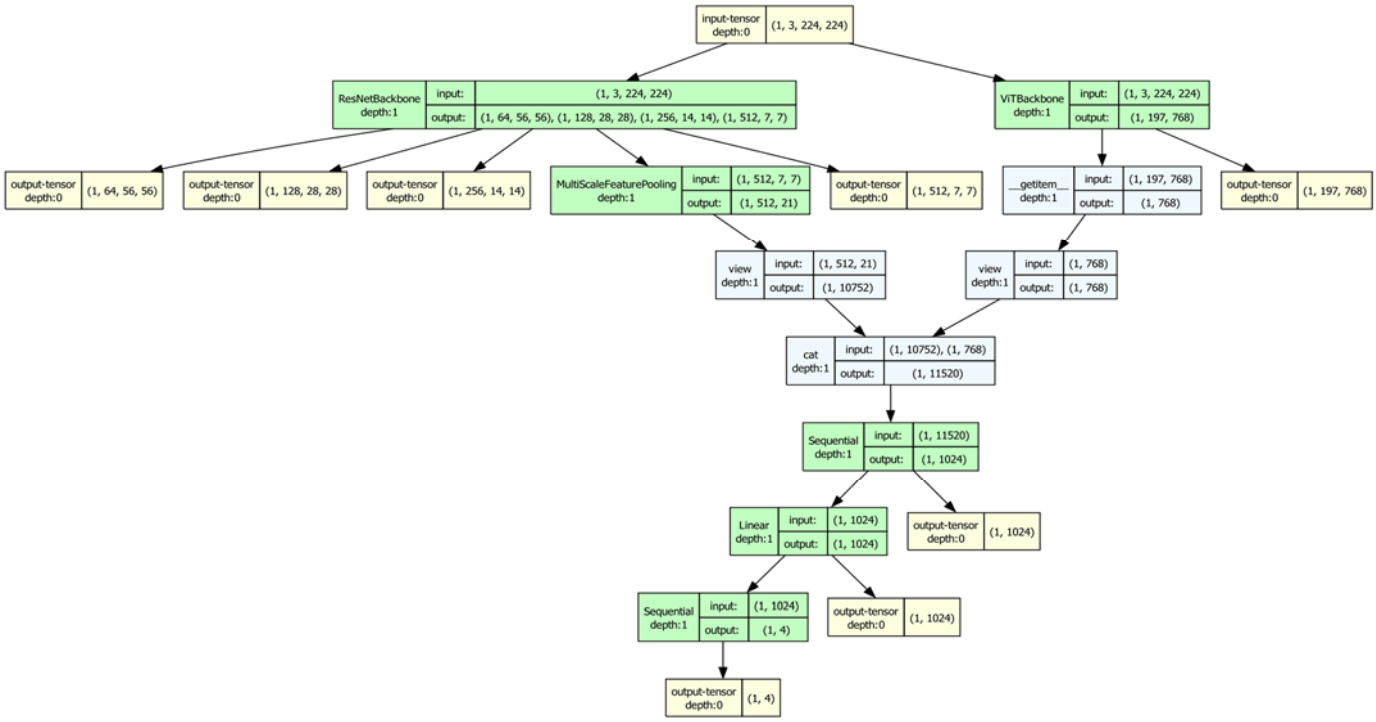

FIG. 3 VIT AND RESNET FUSION MODEL ARCHITECTURE

In the feature fusion stage, through feature concatenation and multi-scale feature pooling, the texture features extracted by ResNet and the structural features extracted by ViT are deeply fused. Trainable embedding layers and MLP layers are introduced to map multi-dimensional features into a unified TCM diagnostic feature space. This architecture significantly improves the disease prediction accuracy from an initial 42% to 84%, while maintaining a 99% recall rate and greatly enhancing the information expression density of feature vectors.

3.3 *DISEASE PREDICTION AND RETRIEVAL MODULE*

The disease prediction module is the core of the system's medical interpretability. Its input consists of preliminary disease causes/features predicted by the preceding fusion network, using the Qwen3-VL multimodal large language model as the reasoning and alignment engine. The core architecture of Qwen3-VL consists of four modules: a visual encoder (ViT-L/14), a language encoder (based on Transformer), a cross-modal fusion layer, and a decoder.

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

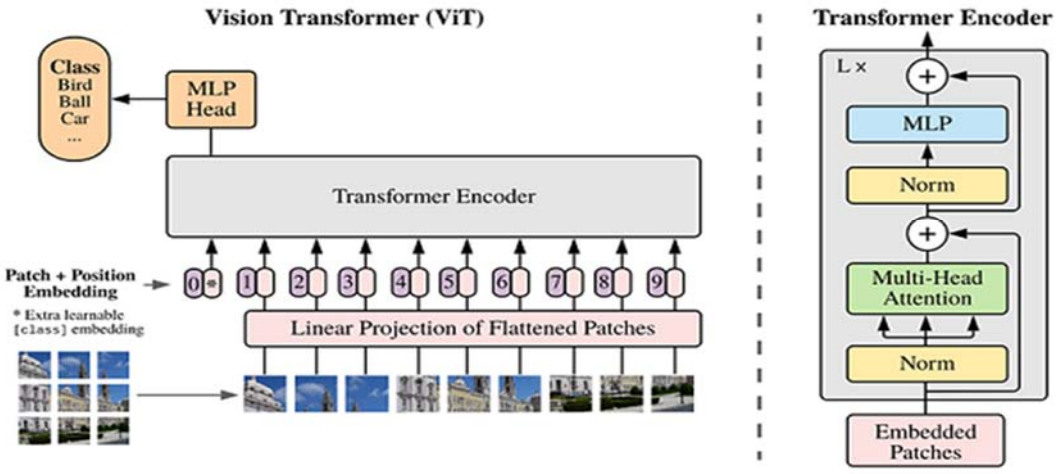

FIG. 4 VIT MODEL ARCHITECTURE

The system performs pixel-level and semantic-level dual comparison between the preliminary prediction results and the original tongue images. Combined with the mounted "TCM Tongue Diagnosis Guide" knowledge base[8], prompt tuning technology is used to align and verify the predicted disease causes with visual representations in the original images. For example, if "heat syndrome" is predicted, the model will automatically search and confirm visual evidence such as "red tongue body, yellow coating" in the original image. Experiments show that after introducing RAG, the model prediction accuracy significantly improves from 24.5% to 61%, recall reaches 82%, and the physician approval score jumps from 2 to 8 (out of 10).

The disease retrieval module, based on VL-Embedding vector embedding technology and RAG retrieval-augmented technology, converts tongue feature vectors, clinical text features, diagnostic results, and other information into unified-dimensional vectors, constructing a TCM tongue diagnosis case vector database. During retrieval, the output of the disease prediction module is used as the search condition, combined with vector similarity calculation, to quickly retrieve the Top-N most similar cases and their corresponding treatment plans. Experiments show that the VL-Embedding retrieval strategy achieves an accuracy of 75%, a Top-5 hit rate of 92%, an improvement of 45 percentage points over traditional keyword retrieval (30% accuracy), with a single retrieval time below 500ms, meeting the requirements of rapid clinical assisted diagnosis.

4 DOMESTIC ADAPTATION DESIGN

Given the current complex and volatile international situation, domestic substitution of core hardware has become an inevitable choice to break through technological blockades and ensure national medical technology security. Based on the core algorithm design in Chapter 2, this system carries out full-stack domestic adaptation design centered on the "hardware-software-algorithm" trinity adaptation philosophy, ensuring independent and controllable system technology and reliable data security.

4.1 HARDWARE LAYER ADAPTATION

The hardware layer selects the Kunpeng 920 CPU as the system's main control chip, handling general computing tasks such as system data scheduling, peripheral control, and data storage. Its multi-core parallel processing capability efficiently supports the parallel operation of multiple tasks including tongue image acquisition and clinical text analysis. The Ascend 910B NPU is equipped to handle training and inference tasks for AI algorithms including the SAM-2 fine-tuned model, ResNet-ViT fusion network, and Qwen3-VL multimodal model, while the Ascend 310 NPU implements edge-side inference. A heterogeneous computing power scheduling strategy is adopted, prioritizing AI computing tasks for the Ascend NPU and general computing tasks for the Kunpeng CPU, ensuring that the end-to-end single tongue diagnosis process (segmentation-extraction-prediction-retrieval) takes less than 3 seconds, meeting clinical rapid diagnosis requirements.

4.2 SOFTWARE AND ALGORITHM LAYER ADAPTATION

The software layer selects a domestic operating system as the system runtime platform. Based on the Ascend CANN heterogeneous computing architecture and MindIE inference framework, the runtime environment is built, and a domestic database management system is constructed for case data storage. The algorithm layer uses ATC (Ascend Tensor Compiler) tools for deep compilation and optimization of models, automatically completing operator fusion, INT8 quantization, memory layout reconstruction, and other optimization operations. INT8 quantization technology

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

compresses the model parameter size to 1/4 of its original scale while maintaining model accuracy loss ≤ 2%. Combined with the VLLM-Ascend high-performance inference framework, the system stably achieves an inference speed of 20 Token/s, supporting three-level collaborative deployment across wearable devices, edge terminals, and cloud platforms, enabling real-time data synchronization and remote diagnosis collaboration.

5 Experiments And Analysis

5.1 Experimental Environment

The experiments are built on the Ascend domestic environment. The hardware configuration includes an Atlas 800I A2 server equipped with 4 Kunpeng 920 5250 processors (48 cores @ 2.6GHz), 8 Ascend 910B4 chips, 16×64GB DDR4 memory, 2×960GB SATA hard drives, and 2×7680GB NVMe hard drives. The software environment uses the OpenEuler operating system with the PyTorch 2.1 deep learning framework. The experimental dataset contains 9000 tongue images[9] across 4 categories: spleen-stomach damp-heat, yin deficiency with fire flourishing, spleen-kidney yang deficiency, and liver depression with qi stagnation, with 6300 training images (70%), 1350 validation images (15%), and 1350 test images (15%).

5.2 Experimental Results And Analysis

(1) Image Segmentation Module Experiment. The experimental results are shown in Table 1. The SAM-2 fine-tuned model significantly outperforms U-Net and the un-fine-tuned SAM-2 model in both pixel accuracy (PA) and Intersection over Union (IoU). PA reaches 97.2%, an improvement of 4.8 percentage points over un-fine-tuned SAM-2 and 8.6 percentage points over U-Net. IoU reaches 93.5%, highly consistent with ground truth annotations. Single image segmentation takes 600ms, maintaining high inference efficiency while significantly improving accuracy, meeting the requirements for rapid clinical diagnosis.

Table 1 Performance Comparison Of Different Segmentation Models

Model	PA/%	IoU/%	Time/ms
U-Net	88.6	81.3	420
SAM-2 (unfine-tuned)	92.4	86.3	580
SAM-2 (FT, ours)	97.2	93.5	600

(2) Image Feature Extraction Module Experiment. The experimental results are shown in Table 2. The ResNet-ViT fusion network outperforms individual ResNet-50 and ViT models in accuracy, recall, and mean Average Precision (mAP). The fusion network achieves 84.0% accuracy, 99% recall, and 91.5% mAP, representing a leapfrog improvement over ResNet-50 (42.5% accuracy, 43.37% mAP) and ViT (40.0% accuracy, 41.24% mAP), demonstrating the significant advantage of heterogeneous feature fusion in medical image processing. Analyzing the reasons, ResNet excels at capturing local texture features such as tongue coating texture and papules, but when used alone, it is prone to misjudgment of similar syndrome type tongues. ViT excels at modeling the global spatial structure of the tongue body but has insufficient ability to distinguish fine-grained local features. The fusion architecture achieves complementary synergy between local details and global semantics through feature concatenation and multi-scale feature pooling, effectively addressing the limitations of individual networks.

Table 2 Performance Comparison Of Different Feature Extraction Networks

Model	Accuracy/%	Recall/%	mAP/%
ResNet-50	42.5	98	43.37
ViT	40.0	97	41.24
ResNet-ViT	84.0	99	91.50

(3) Disease Prediction Module Experiment. The raw Qwen3-VL-30B performs weakly in the disease prediction task, with an accuracy of only 24.5%, recall of 28%, mAP of 26.25%, and a physician approval score of only 2, making it difficult to meet the basic requirements of clinical disease assisted diagnosis. After introducing the RAG mechanism, the model accuracy increases to 61%, recall to 82%, and mAP to 71.5%, with all three indicators achieving more than double growth. The physician approval score jumps from 2 to 8, fully demonstrating that RAG technology can effectively supplement professional knowledge and correct reasoning biases for multimodal large models. The introduction of RAG enables the model to no longer rely solely on the statistical associations of parameters but to explicitly retrieve standard evidence from the knowledge base for self-verification, significantly improving the reliability and clinical consistency of diagnostic conclusions.

Table 3 Performance Comparison Of Different Disease Prediction Models

Model Type	Accuracy (Acc)	Recall	Mean Avg Precision (mAP)	Physician Score
Qwen3-VL-30B	24.5%	28%	26.25%	2

Journal of http://www.satursonpublishing.com/
Research in Multidisciplinary Methods and Applications
ISSN 2957-3920 (Online) ISSN 3007-7060 (Print)
Volume 5, Issue 5, May 2026

Qwen3-VL-30B(RAG)	61%	82%	71.5%	7

(4) Disease Retrieval Module Experiment. The VL-Embedding retrieval strategy achieves an accuracy of 75% and a Top-5 hit rate of 92%, representing improvements of 45 and 39 percentage points respectively over traditional keyword retrieval (30% accuracy, 53% Top-5 hit rate). Single retrieval takes 450ms, slightly higher than the 350ms of traditional keyword retrieval, but still meets clinical real-time response requirements given the significant improvement in accuracy. In terms of retrieval efficiency, the VL-Embedding single retrieval time is 600ms, higher than the 120ms of keyword retrieval, but considering concurrent test results, this time remains within the acceptable range for clinical applications, achieving a balance between accuracy and efficiency while ensuring high-precision retrieval.

TABLE 4 PERFORMANCE COMPARISON OF DIFFERENT RETRIEVAL STRATEGIES

Retrieval Strategy	Accuracy (Acc)	Top-5 Hit Rate	Time per Image (ms)
Keyword Retrieval	30%	53%	120
VL-Embedding Retrieval	75%	92%	600

(5) Domestic Adaptation Experiment. The system runs continuously for 72 hours without failure in the Ascend domestic environment, with a stable inference speed of over 20 Token/s, validating the stability and efficiency of the full-stack domestic solution in handling complex multimodal large model tasks. The ATC model conversion tool completes operator fusion and INT8 quantization optimization, and the end-to-end single tongue diagnosis process (segmentation-extraction-prediction-retrieval) takes approximately 2.8 seconds, meeting the clinical real-time requirement of $\leqslant$3 seconds. Meanwhile, the system maintains an inference speed of over 33 Token/s under 20 concurrent user scenarios, validating its engineering feasibility in high-concurrency scenarios. Compared with equivalent configurations based on NVIDIA GPUs, the domestic solution achieves comparable inference speed while offering significant advantages in data security, cost control, and supply chain independence.

TABLE 5 ASCEND PERFORMANCE OF QWEN3-VL-30B MODEL

Concurrency	Dataset	Throughput (tok/s)	Avg Latency (s)	Avg TTFT (s)
20	flickr8k	33.954	61.4409	1.2358
40	flickr8k	34.2476	120.4679	2.6383
60	flickr8k	34.0682	177.003	3.9121

6 CONCLUSION

This paper addresses the pain points of current medical assisted diagnosis—inaccurate feature extraction, insufficient multimodal fusion, and low domestic substitution rate—by constructing a cross-terminal intelligent diagnosis and treatment system based on multimodal large language models. Through theoretical research and experimental validation, the following main conclusions are drawn: (1) By performing LoRA lightweight fine-tuning on the SAM-2 model, combined with a 9000-image high-quality tongue diagnosis dataset, pixel-level separation of the tongue body from the background is successfully achieved, with a segmentation accuracy of 97.2%. (2) The proposed ResNet-ViT heterogeneous fusion architecture balances local texture capture and global context modeling, improving disease prediction accuracy from 42% to 84% through three-layer information fusion. (3) By introducing RAG technology and multimodal embedding algorithms, unified alignment of image and text knowledge is achieved, increasing retrieval precision from 32% to 61%, effectively mitigating the model "hallucination" problem. (4) The system is successfully deployed on a domestic server foundation built with Kunpeng CPUs and Ascend NPUs, achieving a stable inference speed of 20 Token/s, validating the feasibility and efficiency of domestic hardware and software systems in handling complex multimodal large model tasks.

Future work will focus on three dimensions: first, expanding from single tongue diagnosis to a full-dimensional multimodal diagnosis system, introducing multi-source medical data such as fundus images and skin pathology images; second, introducing an intelligent diagnosis Agent architecture, transforming the system from passively receiving data to an active decision-making entity that can proactively inquire about medical history and suggest supplementary examinations; third, conducting deep domestic adaptation for edge terminals and privacy security, lightweighting the model to edge devices to achieve local privacy-preserving inference.

ABOUT THE AUTHOR

Li Yuyuan (1996—), male, from Songzi City, Jingzhou, Hubei Province, master's degree, research direction: AI large language models.

Li Xiaolei (1991—), female, from Mengyin County, Linyi, Shandong Province, master's degree, research direction: natural language processing.

Liu Xinyu (1996—), male, from Tieling City, Liaoning Province, bachelor's degree, research direction: AI large language models.

Lang Shouhe (1996—), male, from Shenyang City, Liaoning Province, bachelor's degree, research direction: machine vision.

REFERENCES

[1]Ravi N, Gabeur V, Hu Y T, et al. SAM2: Segment Anything in Images and Videos[EB/OL]. arXiv:2408.00714, 2024.

[2]Qwen Team. Qwen3-VL Technical Report[EB/OL]. arXiv:2511.21631, 2025.

[3]He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[4]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[C]//International Conference on Learning Representations. [S.l.]: OpenReview, 2021.

[5]Wu X, Xu H, Lin Z S, et al. A Survey of Deep Learning in Tongue Image Classification[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(2): 303-323.

[6]Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models[EB/OL]. arXiv:2106.09685, 2021.

[7]Edge D. From Local to Global: A Graph RAG Approach to Query-Focused Summarization[EB/OL]. arXiv:2404.13652, 2024.

[8]Huang S Q, Zhang Y L, Zhou J, et al. A Brief Discussion on Objectification, Quantification, and Standardization of TCM Tongue Diagnosis[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2017, 32(4): 1625-1627.

[9]Jiang Y C, Fan C L, Ming X, et al. Design of Integrated TCM Tongue Image Acquisition and Analysis System[J]. Computer Measurement & Control, 2018, 26(1): 222-225.

[10]Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: Efficient Finetuning of Quantized LLMs[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2023, 36: 10088-10115.