Original Paper

Application of a Large Visual Language Model on Tongue Image Description Generation and Physical Constitution Reasoning in Traditional Chinese Medicine (TongueVLM): Model Development and Validation Study

Chengdong Peng[1,2], MSc, PhD; Jun Gao[1], Prof Dr, PhD; Nuo Yang[2,3], BE; Yong Wang[2,3], BE; Renming Chen[2], MSc; Changwu Dong[4], Prof Dr Med

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui, China
[2]Artificial Intelligence Laboratory, Hefei Yunzhen Information Technology Co, Ltd, Hefei, Anhui, China
[3]School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei, Anhui, China
[4]The Second Clinical Medical School, Anhui University of Traditional Chinese Medicine, Hefei, Anhui, China

Corresponding Author:
Chengdong Peng, MSc, PhD
School of Computer Science and Information Engineering
Hefei University of Technology
Number 193 Tunxi Road
Hefei, Anhui, 230009
China
Phone: 86 055165337378 ext 86
Email: 2020020018@mail.hfut.edu.cn

Abstract

Background: In the field of traditional Chinese medicine (TCM), diagnostic work based on tongue images to recognize the physical constitution is a process of collecting clinical information, reasoning, and combining the patient's tongue image features with questioning. It is necessary to simulate the recognition of pathological information of tongue images by TCM practitioners and professional dialogue based on tongue image features, which helps to develop an intelligent interactive system for TCM diagnosis.

Objective: This study aimed to develop and validate a vertical model of the TCM domain with TCM's understanding and reasoning capability for tongue images.

Methods: A TongueVLM multimodal large model is designed, which includes a visual encoder module, a modal fusion module, and a language decoder module. First, the visual encoder based on the CLIP-ViT (Contrastive Language-Image Pre-Training With Vision Transformer) pretrained model is used for image patch, dimensionality reduction, and migration learning, which maps the high-dimensional tongue features into low-dimensional language encoding vectors. Further, a modal fusion module with a residual architecture is applied to map visual features to a natural language word embedding space, realizing the conceptual alignment between visual encoding and TCM terminology. Finally, fine-tuning of visual instructions is performed based on the LLaMA (large language model meta artificial intelligence), and a TCM-domain large language model with 7B parameters is trained.

Results: The constructed multimodal dataset has 3 test datasets, and experiments are conducted using 3000 samples from each test dataset, respectively. Experimental results indicate that the TongueVLM model outperforms general-purpose large models on all 3 tasks. On the multimodal test dataset, the TongueVLM model achieved accuracy rates of 79.8%, 78.6%, and 60.7% in evaluation tasks respectively, it achieves 9.1%, 8.4%, and 1.1% in greater accuracy than LLaVA-OneVision, and is 7.5%, 7%, and 5.9% more accurate than Qwen2.5-VL-7B, with the text generation time being around 24 tokens per second.

Conclusions: The TongueVLM model, which achieves tongue image description generation and physical constitution reasoning in TCM, is suitable for the application of a Chinese medicine intelligent diagnosis system.

(JMIR Med Inform 2026;14:e87237) doi: 10.2196/87237

KEYWORDS

vision-language model; tongue image description generation; visual constitution reasoning; visual dialogue; intelligent diagnosis of traditional Chinese medicine (TCM)

Introduction

In recent years, the research focus on computer vision has gradually shifted toward innovative transformers and their variant architectures. These studies have made significant progress in the 3 basic tasks of image classification, object detection, and image segmentation, along with in-depth experiments on visual-textual multimodal data validity. The visual transformer model, represented by the vision transformer, approaches or even surpasses the convolutional neural network approach in several benchmark tests [1-14]. The key technique of the vision transformer is to construct image-to-vector transformations and effectively maintain the features of the image, bridging the gap between language and vision. Lava family [15-17], GPT-4v [18], and Gemini 1.5 [19] visual-language multimodal modeling have even ushered artificial intelligence-generated content into a new era of visual applications, and the field of medicine has received increasing attention from researchers [20-26].

In the field of traditional Chinese medicine (TCM), diagnostic work based on tongue images to recognize the physical constitution is a process of collecting clinical information, reasoning, and combining the patient's tongue image features with questioning. It is necessary to simulate the recognition of pathological information of tongue images by TCM practitioners and professional dialogue based on tongue image features, which helps to develop an intelligent interactive system for TCM diagnosis. However, the following limitations still exist in the current investigation: (1) high cost of acquisition equipment and collection of clinical data, mostly relying on publicly available datasets, and prominent data sample noise problems; (2) lack of qualitative or quantitative specialized labeling data; (3) lack of validation and analysis of the visual encoding results at the modal fusion stage; and (4) the evaluation metrics of the trained models focus on technical performance and ignore professional human judgments.

Previous studies have shown that, based on the image recognition model in the field of TCM applications, they fused the traditional image features of the tongue with deep tongue features and combined them with a machine learning model to construct a physical constitution identification, which compensated for the lack of information in a single modality and improved the accuracy of constitution classification [27-29]. In terms of training a vertical model of TCM based on a large language model (LLM), we constructed a pretraining dataset and an instruction fine-tuning dataset oriented to TCM and adopted a 2-phase training method of continuous pretraining and supervised fine-tuning to develop an LLM of TCM [30-35]. This model outperforms the existing models in several evaluation metrics and can play a role in TCM consultation.

Medical applications based on multimodal large models are also favored by researchers. A dataset containing 5.3 million+ image-text pairs was used to train the UniMed-CLIP model, which outperforms the existing health care visual language models (VLM) with 0-shot performance [36]. Multimodal health care models, which introduce evaluation metrics and frameworks and perform well in multimodal Q&A (question and answer) and text-based Q&A tasks [33,37]. BiomedCLIP with LLaMA-3 (large language model meta artificial intelligence) to realize a lightweight and efficient medical visual Q&A model [38]. Med-VQA interpretability by localizing image preference regions [39] enhanced the performance of biomedical VLMs by using instruction fine-tuning and multi-image training, respectively [40,41]. Aligned visual features with medical concepts and Med-VQA, enhanced the generation of large models, and achieved excellent performance on public datasets [42-44].

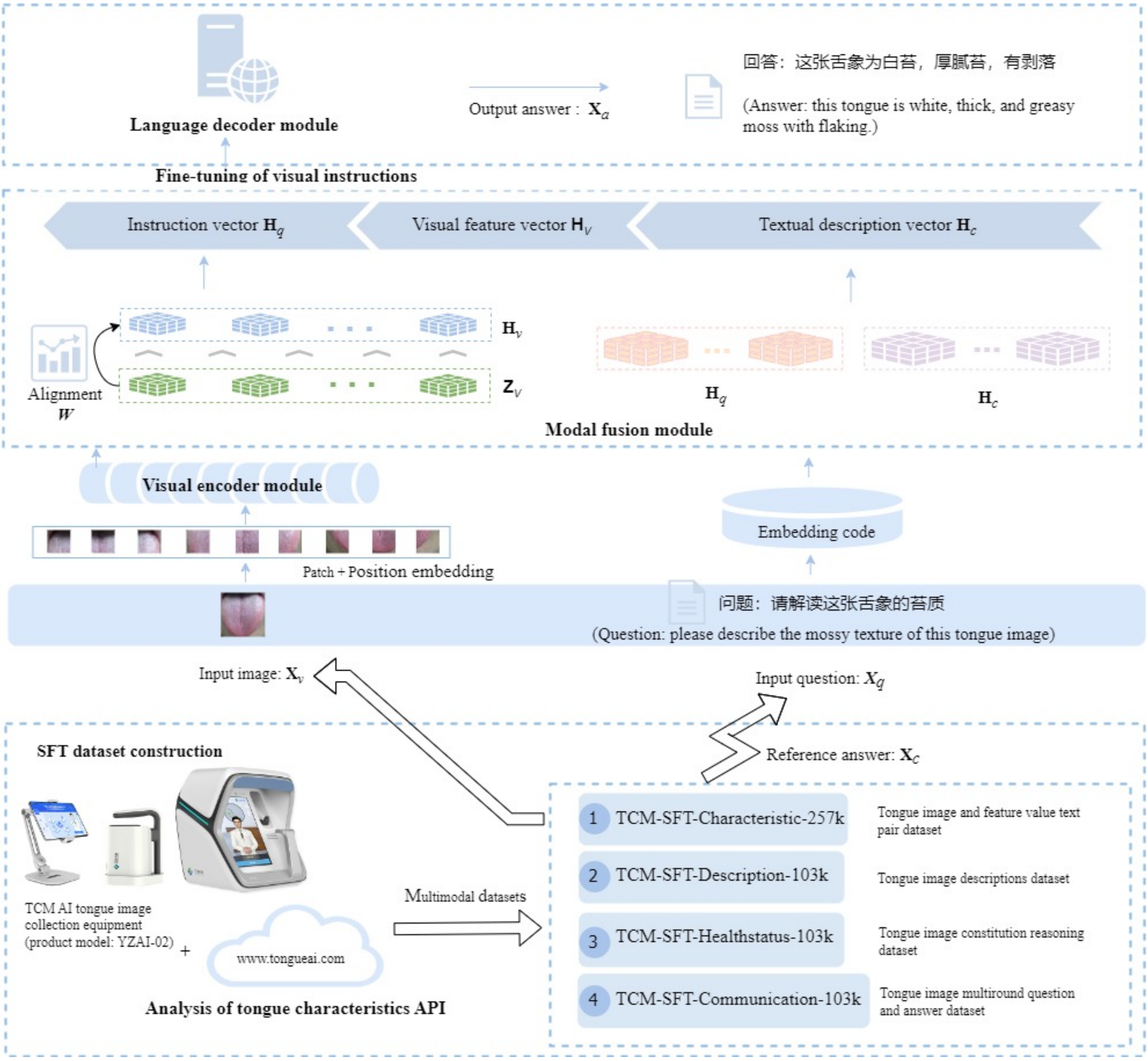
The primary aim of this study is to develop and validate TongueVLM, a specialized VLM for TCM tongue diagnosis. Specifically, our objectives were to (1) design a novel multimodal architecture that effectively aligns visual features of tongue images with domain-specific TCM textual knowledge; (2) construct a high-quality, multimodal dataset for supervised study, and end-to-end training on the TCM dataset; and (3) evaluate the performance of TongueVLM against established baseline models, and validation multimodal capabilities across 3 tasks of tongue feature description generation, physical constitution reasoning, and multiround dialogue.

Methods

Overall Architecture of TongueVLM

Our TongueVLM model uses a hybrid architecture combining the visual encoder module, modal fusion module, and language decoder module. First, the visual encoder based on the CLIP-ViT (Contrastive Language-Image Pre-Training With Vision Transformer) pretrained model is used for image patch, dimensionality reduction, and migration learning. Further, a multimodal fusion module with a residual architecture is applied to map visual features to the natural language word embedding space, realizing the conceptual alignment between visual encoding and TCM terminology. Finally, a visual-textual instruction dataset is constructed to perform multimodal instruction fine-tuning based on the LLaMA general-purpose LLM. As a result, a vertical model of the TCM domain with TCM understanding and reasoning capability for tongue images is obtained. The realization procedure is shown in Figure 1.

Figure 1. Network structure and computation of a large visual-language model. Includes a visual encoder module, a modal fusion module, and a language decoder module. First, the visual encoder based on the CLIP-ViT pretrained model is used for image patch, dimensionality reduction, and migration learning. Further, a modal fusion module with a residual architecture is applied to map visual features to a natural language word embedding space. Finally, fine-tuning of visual instructions is performed based on the LLaMA. AI: artificial intelligence; API: application programming interface; CLIP ViT: Contrastive Language Image Pre Training With Vision Transformer; LLaMa: large language model meta artificial intelligence; SFT: supervised fine-tuning; TCM: traditional Chinese medicine.


Language decoder module
Output answer : $\mathbf{X}_a$
回答：这张舌象为白苔，厚腻苔，有剥落
(Answer: this tongue is white, thick, and greasy moss with flaking.)
Fine-tuning of visual instructions
Instruction vector $\mathbf{H}_q$
Visual feature vector $\mathbf{H}_v$
Textual description vector $\mathbf{H}_c$
Alignment W
$\mathbf{H}_v$
$\mathbf{Z}_v$
$\mathbf{H}_q$
$\mathbf{H}_c$
Modal fusion module
Visual encoder module
Embedding code
Patch + Position embedding
问题：请解读这张舌象的苔质
(Question: please describe the mossy texture of this tongue image)
Input image: $\mathbf{X}_v$
Input question: X_q
Reference answer: $\mathbf{X}_c$
SFT dataset construction
TCM AI tongue image collection equipment (product model: YZAI-02)
+
www.tongueai.com
Analysis of tongue characteristics API
Multimodal datasets
1 TCM-SFT-Characteristic-257k
Tongue image and feature value text pair dataset
2 TCM-SFT-Description-103k
Tongue image descriptions dataset
3 TCM-SFT-Healthstatus-103k
Tongue image constitution reasoning dataset
4 TCM-SFT-Communication-103k
Tongue image multiround question and answer dataset

Model Design

Phase 1: Visual Encoder Module

In the first phase, the visual encoder module of the network was designed. Applying tongue images with TCM feature description text pair data in CLIP-ViT pretrained model migration learning. Semantically rich text is used as training labels to project the image feature representations into a kind of encoding that is like the text encoding space, as shown in Figure S1 in Multimedia Appendix 1.

First, we speak of image patch dimensionality reduction preprocessing. Scale the input image $\mathbf{x}_v\in(H,W,C)$ to $\mathbf{x}\in(h,w,C)$ (h is the height, w is the width, and C is the number of channels). Sequence of blocks that transform x into $\mathbf{x}_p\in(N^2,p^2C)$ for image patch operations. The number of blocks is $N^2=(\mathrm{h}\times w)/p^2$, and the block dimension is (p^2C); where p is the image patch size. A sequence of blocks is converted into a patch embedding vector $\mathbf{E}_{patch}{}'\in(N^2,dim)$ by a linear transformation, where dim=1024, p=14, and N= 24. $\mathbf{x}_0$ is the zeroth encoding vector taken at the output layer of the transformer encoder, which serves as the visual feature encoding. By adding a learnable embedding vector before the image patch embedding vector $\mathrm{E}_{\mathrm{patch}}$, the image block embedding vector becomes $\mathbf{E}_{patch}\in(N^2+1,dim)$. The absolute position encoding algorithm is used to create a learnable position embedding vector $\mathbf{E}_{pos}\in(N^2+1,dim)$ for the block position. The elementwise summation operation on $\mathbf{E}_{patch}$ and $\mathbf{E}_{pos}$ yields an embedding vector $\mathbf{z}\in(N^2+1,\mathrm{dim})$ that incorporates spatial positional encoding and visual representations.

Second, transformer encoder processing. The encoder consists of a stack of $\mathbf{L}$=24 encoding layers, and the encoder inputs are image feature embedding vectors $z\in(N^2+1,dim)$. Each encoding layer consists of 2 unitary structures connected: the first unitary structure consists of a multihead self-attention layer and a layer normalization, as well as a residual connectivity module, and the second unitary structure consists of a feedforward fully connected layer multilayer perceptron and a layer normalization, as well as a residual connectivity module.

Phase 2: Modal Fusion Module

The modal fusion module is designed with a multilayer linearly transformed residual architecture, which is a bridging module between the visual model and the language generation model. The tongue visual encoding features are projected into the natural language word embedding vector feature space to solve the conceptual alignment problem of visual encoding with TCM terminology.

The modal fusion module design uses a combination of multiple linear layers and GELU (Gaussian Error Linear Unit) activation functions, as shown in the middle region of Figure S2 in Multimedia Appendix 1. This design allows the modal fusion layers to transform visual features while maintaining essential nonlinear characteristics. In addition, the introduction of the residual structure enables the model to realize efficient feature transformation and projection while fully preserving the integrity of the original visual features.

Splice visual encoding with textual word-embedded encoding to form a sequence of visual-textual instructions. First, the visual encoding feature $\mathbf{Z}_v$ is converted into a visual feature vector $\mathbf{H}_v$, and the textual instruction and textual description are converted into an instruction vector $\mathbf{H}_q$ and a textual description vector $\mathbf{H}_c$ by word embedding encoding, respectively. Thereafter, the vectors are spliced in the order of instruction, visual, and textual description to form the visual-textual fusion instruction sequence data.

Phase 3: Language Decoder Module

The modal fusion module connects the visual encoder and language decoder for end-to-end fine-tuned training on the constructed visual-textual data for a variety of downstream generative output tasks, such as visual quizzing, image description generation, and visual dialogue. The language decoder module chooses LLaMA, a pretrained model with good support for Chinese, which consists of a 32-layer transformer decoder layer, and each decoder layer contains an RMSNorm normalization layer, a grouping attention mechanism, and a multilayer perceptron projection layer. The network architecture is shown on the right side of Figure S2 in Multimedia Appendix 1.

During training, each input image $\mathbf{X}_v$ is associated with a few rounds of dialogue data, where each round of question description $X^t_{instruct}$ can be defined as follows.

$$\mathbf{X}^t_{instruct}=\begin{cases}[(\mathbf{X}_v,\mathbf{X}^1_\mathbf{q}),\mathbf{X}^1_c],\ t=1\\ (\mathbf{X}^t_q,\mathbf{X}^t_c),\ t>1\end{cases}\quad(1)$$

The visual encoder module extracts the visual feature V_x (the feature representation obtained by encoding the input image using a visual encoder) and caches it into the context, and the subsequent t-th round (t>1) of training data only needs to input the new instruction $\mathbf{X}^t_q$ and answer $\mathbf{X}^t_c$. The model automatically fuses the historical dialogue information with the initial visual feature $\mathbf{V}_x$ and builds a spliced sequence of instructions $X^t_{insruct}$ to realize multiple rounds of dialogue training through cross-round dependency.

The instruction sequence $\mathbf{X}_{instruct}=\{\mathbf{x}^1_{instruct}, \mathbf{x}^2_{instruct}, \dots, \mathbf{x}^t_{instruct}\}$ of multiple rounds and the corresponding output reply text $\mathbf{X}_a=\{\mathbf{x}^1_a,\mathbf{x}^2_a,\dots,\mathbf{x}^\mathrm{t}_a\}$ are temporally spliced into a uniform sequence $\{X^1_{insruct}\mathbf{x}^1_a, X^2_{insruct}\mathbf{x}^2_a, \dots ,X^t_{insruct}\mathbf{x}^\mathrm{t}_a\}$, the autoregressive training objective is used for instruction tuning, and the maximum likelihood function optimization model is formulated as follows:

$$p(\mathbf{X}_a|\mathbf{X}_v,\mathbf{X}_{instruct})=\prod_{i=1}^{\mathbf{L}}p_\theta\ (x_i|\mathbf{X}_v,\mathbf{X}_{instruct,<i},\mathbf{X}_{\mathrm{a},<i})\ (2)$$

where θ is a trainable parameter (the training process is to adjust to maximize the likelihood function) and $\mathbf{X}_{instruct,<i},\mathbf{X}_{a,<i}$ is the instruction sequence and answer sequence of all rounds before the current prediction x_i, respectively. $p(\mathbf{X}_a|\mathbf{X}_v,\mathbf{X}_{instruct})$ is the probability of predicting $\mathbf{X}_a$ under the current $\mathbf{X}_v$ and $\mathbf{X}_{instruct}$ is based on the premise of guaranteeing the accuracy of the result of $\mathbf{X}_a$under the premise that the prediction results of all rounds before the current round takes the maximum probability value.

In other words, the model uses the information accumulated in the history of the previous *t*-1 rounds of the conversation (especially the context of the textual instructions) to accurately estimate the probability distribution of the generated results in the t-th round. In this way, the model can maintain its understanding of the context over consecutive rounds of interactions and accordingly generate accurate responses associated with the content of the previous rounds.

Datasets and Preprocessing

The 90,000 tongue images collected by the TCM artificial intelligence tongue image collection equipment (product model: YZAI-02) jointly developed by Hefei Cloud Diagnostics Information Technology Co and Anhui University of TCM were used, which consists of a standard LED ring light source, a portable collection device, an ultrahigh-definition camera, and a host computer, as well as supporting software. The tongue feature recognition API (application programming interface) was applied to preprocess the collected tongue images with color correction, tongue region segmentation, and tongue moss and tongue texture separation, after which tongue feature analysis was performed on the tongue images to obtain more than 40 types of qualitative or quantitative text containing tongue color, tongue texture, moss color, moss texture, and sublingual collateral data and 10 types of TCM constitution [45].

The TongueVLM training datasets are constructed, including TCM feature description data (traditional Chinese medicine–supervised fine-tuning, TCM-SFT (TCM–supervised fine-tuning)-Characteristic-257k, and TCM-SFT-Description-103k), multiround Q&A data (TCM-SFT-Healthstatus-103k), and physical constitution reasoning data (TCM-SFT-Communication-103k). The methodology for constructing the dataset is described in Multimedia Appendix 1. The distribution of the tongue image features in the dataset is shown in Figures S3 and S4 in Multimedia Appendix 1.

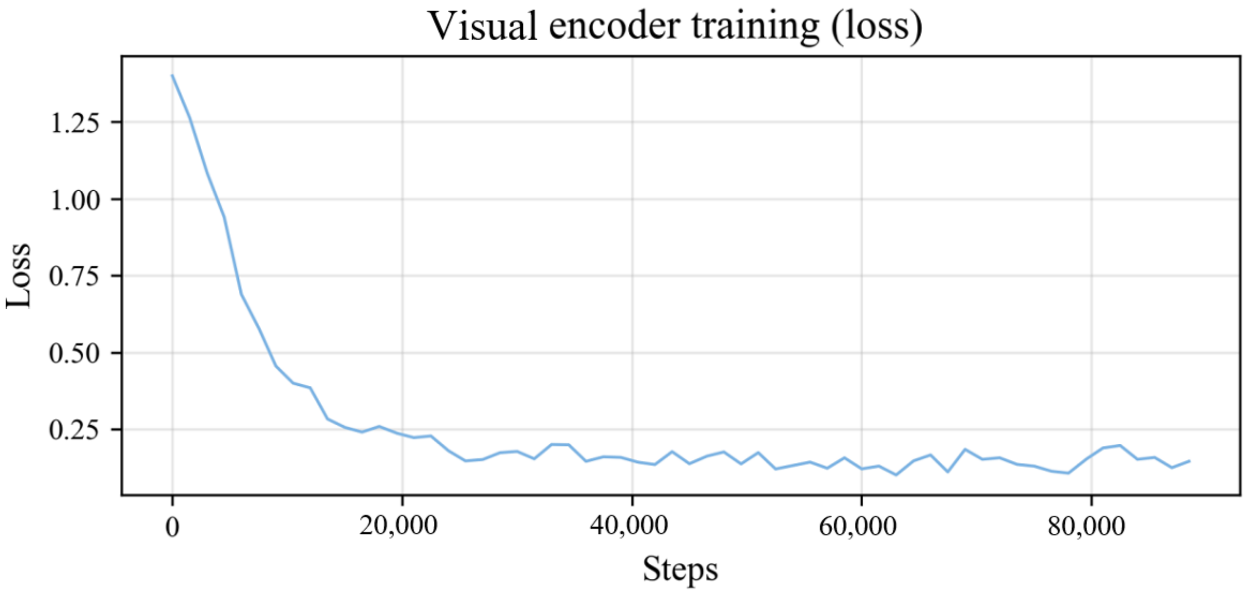
Model Training Strategy

Stage 1: Pretraining of the Visual Encoder

TCM tongue characterization datasets (TCM-SFT-Characteristic-257k) of 257,000 tongue images and characterization texts are constructed, and the zeroth encoding vector used for learning by comparing the similarity between the tongue images and their corresponding tongue manifestation characterization is used for training the fine-tuning of the tongue visual encoder.

The process to obtain the visual encoder model is as follows: first, the training parameters openai/clip-vit-large-patch14-336 (OpenAI Inc, pretrained weights for the clip model on 400 million text-image pairs) were used as the initial weights for the visual encoder and the text transformer encoder. The Adam optimizer was used in the model training, with the smoothing constant betas set to (0.9, 0.999) and the weight decay parameter weight_decay set to 1e-3, the training batch size was set to 8, the learning rate was 5e-5, the number of training epochs was 3, and the maximum sequence length was 77. Second, the loss function uses the cosine similarity $\mathrm{Cos}(I_m,T_n),m\in[1\ldots N^2],n\in[1\ldots N^2]$, where I_m is the embedding vector encoded by visual features and T_n is the embedding vector encoded by textual semantics. The optimization objective is to maximize the cosine similarity $\mathrm{Cos}(I_j,T_j)$ between correct image-text pair embeddings while minimizing the cosine similarity $\mathrm{Cos}(I^j,T^k),j\neq k$ of incorrect image-text pair embeddings. The results show that the loss value stabilizes after 3 epochs of training, and the loss curves are shown in Figure 2.

Figure 2. Visual encoder training loss and dynamic learning rate curves. The loss curve shows that after 3 rounds of training, the training loss of modal fusion has gradually stabilized, indicating that the model is beginning to converge.

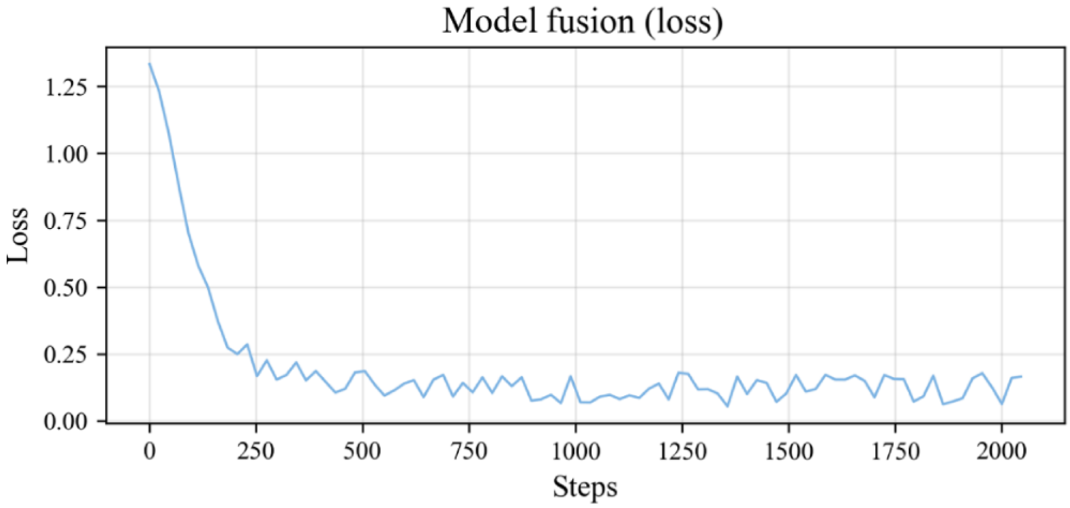


Stage 2: Training of the Fusion Modal

TCM tongue description datasets (TCM-SFT-Description-103k) of 103,000 tongue images and description text are constructed, the visual encoder of TongueVLM and the network weights of the LLM are frozen, the visual-textual multimodal instruction data are fed into the LLM, and the modal fusion module is fine-tuned for training.

The process to obtain the modal fusion model is, first, the training method and parameters are as follows: the pretrained visual encoder weight parameters are loaded, the modal fusion layer weight parameters are randomly initialized, and only the modal fusion layer weights are allowed to participate in the gradient update. The Adam optimizer was used in the model training, with the smoothing constant betas set to (0.9, 0.999) and the weight decay parameter weight_decay set to 1e-4, the training batch size on each GPU is 8, the gradient accumulation step is 8, the learning rate is 2e-5, and the total number of iteration rounds is 3. Spend 30 hours to complete all the training on the Ubuntu (version 20.04.6, Canonical Ltd) platform using 2×RTX4090. Second, the loss function adopts the cross-entropy loss $\mathbf{Loss}=-\frac{1}{L}\cdot\sum_{i=1}^{L}\sum_{n=1}^{C}y_{n,i}*\log S_{n,i}$, where C is the total number of word lists, L is the length of the prediction sequence, $S_{n,i}$ denotes the probability that the word in the ith position is the nth word in the lexicon based on the previously known sequence, and $y_{n,i}$ denotes the ground-truth labels in the ith position in the prediction sequence (expressed as 1-hot encoding; ie, if the corresponding word in the position is the nth vocabulary word in the lexicon, only the first element of the 1-hot vector is 1 and the rest are 0), the goal is to minimize the loss; when the probability of each word in the prediction sequence is infinitesimally close to the actual vocabulary 1-hot encoding, that is, for any position *i* in the prediction sequence, $S_{n,i} \approx y_{n,i}$, the loss will be minimized, and at this time, the model is considered to be in the optimal state. The results show that the loss value stabilizes after 3 epochs of training, and the loss curves are shown in Figure 3.

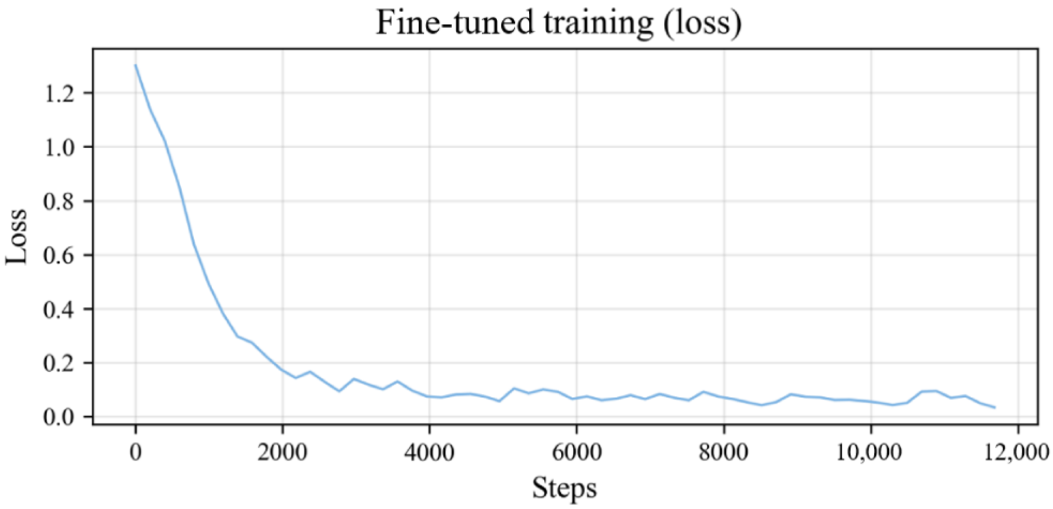
Figure 3. Loss curve of the modal fusion model. It shows that after 3 training epochs, the model enters a stable convergence state, and the loss curve for its modal fusion training stabilizes.



Stage 3: End-to-End Fine-Tuning of TongueVLM

The training data are the TCM multimodal dataset (training set 510,000; validation set 25,000). First, of the training method and parameters, load the pretrained visual encoder weight parameters and modal fusion layer weight parameters, freeze the weight parameters of the visual encoder, and allow only the modal fusion layer and the language decoder weights to participate in the gradient update. The Adam optimizer was used in the model training, with the smoothing constant betas set to (0.9, 0.999) and the weight decay parameter weight_decay set to 1e-4, the training batch size is set to 8 on each GPU, the number of gradient accumulation steps is 8, the learning rate is 2e-5, and the total number of iteration rounds is 3. It takes 135 hours to complete all the training on the Ubuntu (version 20.04.6, Canonical Ltd.) platform using 2×RTX4090. Second, of loss function, the cross-entropy loss function is adopted, and the loss curve is shown in Figure 4. After 4000 iterations, the loss tends to stabilize, and the loss fluctuation is very small when the number of iterations is 8000; currently, the model has converged.

Figure 4. Fine-tuned training loss curve of TongueVLM. It shows that with the increase of training epochs, the fine-tuning training loss of TongueVLM gradually tends to stabilize, indicating the effectiveness of the optimization algorithm and the basic convergence of the model after about 8000 iterations.



Ethical Considerations

This study was conducted in accordance with the ethical standards of the Declaration of Helsinki and was approved by the ethics committee of the Second Clinical Medical School of Anhui University of Chinese Medicine (Anhui Acupuncture and Hospital). The data used in this research were sourced from 2 distinct origins. First, for data from the Yunzhen 365 app, the data consisted of deidentified user-uploaded tongue images and associated text. All users of the app provided informed consent upon registration, agreeing to the use of their anonymized data for research purposes in accordance with the app's terms of service and privacy policy. Second, for data from hospital records, the clinical data were obtained with approval from the relevant hospital's institutional review board (2022-zjks-25). Patient confidentiality was strictly maintained, and all personal identifiers were removed before analysis.

All procedures involving human data were reviewed and approved by the appropriate ethics committees. No additional identifiable personal information was accessed or used in this study. Participants received no compensation for their involvement in this study.

Results

Overall Performance of TongueVLM on Benchmark Tasks

The performance evaluation of the TongueVLM on the TCM multimodal test dataset is as follows.

To determine whether the answer texts generated by professional TCM practitioners are close in meaning to the standard answer texts in the corresponding test set, the evaluation is divided into 3 results; according to the degree of similarity in meaning, we define a score value Score for it, in descending order: consistent (Score=100), similar (Score=50), and irrelevant (Score=0); the accuracy of TongueVLM on the test set can be expressed as follows: $\mathrm{Acc}=\frac{1}{N}\cdot\sum_{i=1}^{N}\mathrm{Score}_i$, where N is the number of test datasets and Score_i is the score of the i-th test dataset.

To directly verify the effectiveness of the improved methods proposed in this study for the field of TCM, we selected visual multimodal models with similar architectures for comparison: LLaVA-1.5 (released in October 2023; Meta Platforms, Inc), LLaVA-1.6 (released in January 2024; Meta Platforms, Inc), and LLaVA-OneVision (released in August 2024; Meta Platforms, Inc). To demonstrate that good results can be achieved in the field of TCM through targeted model design without relying on massive general data, we also compared it with the current open-source multimodal large model benchmark Qwen2.5-VL-7B (released in August 2024; Alibaba Cloud). The accuracy and performance are shown in Table 1. The experimental results show that the TongueVLM model outperforms the general-purpose large model on all 3 tasks; it is 9.1%, 8.4%, and 1.1% more accurate than LLaVA-OneVision and 7.5%, 7%, and 5.9% more accurate than Qwen2.5-VL-7B. The generation time efficiency of each large model is similar, with the generation time of text being 24 tokens per second (approximately 40 words per second, given the word-to-token ratio in our dataset). Due to the optimization of the lexicon in the field of TCM, TongueVLM inference speed is slightly faster than the baseline model.

Table 1. Evaluation indicators for the output of the model. Shows the evaluation metrics of the model output. Experiments are conducted using the test dataset of 3 multimodal datasets, with each test dataset containing 3000 samples with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets.

Model	TCM-SFT[a]-Description-103k			TCM-SFT-Healthstatus-103k			TCM-SFT-Communication-103k		
	Test result details	Accuracy (%)	Mean time (SD, s)	Test result details	Accuracy (%)	Mean time (SD,s)	Test result details	Accuracy (%)	Mean time (SD,s)
CNN[b] parallel computing	167,58,2775/3000	79	1.48 (>6)	—[c]	—	—	—	—	—
LLaVA-1.5	1730,973,297/3000	26.1	0.27 (1.25)	1828,920,252/3000	23.7	0.24 (1.5)	1174,1653,173/3000	33.3	0.64 (3)
LLaVA-1.6	568,1285,1147/3000	59.7	0.21 (1.55)	545,1241,1214/3000	61.2	0.21 (1.5)	690,1094,1216/3000	58.8	0.62 (3)
LLaVA-OneVision	393,974,1633/3000	70.7	0.17 (1.54)	423,941,1636/3000	70.2	0.25 (1.5)	635,1156,1209/3000	59.6	0.86 (3)
Qwen2.5-VL-7B	412,836,1752/3000	72.3	0.18 (1.48)	420,866,1714/3000	71.6	0.15 (1.4)	519,1672,809/3000	54.8	0.69 (3)
TongueVLM	305,599,2096/3000	79.8	0.11 (1.25)	353,577,2070/3000	78.6	0.20 (1.5)	428,1502,1070/3000	60.7	0.54 (3)

[a] TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

[b] CNN: convolutional neural network.

[c] Not available.

Ablation Studies on the 3-Stage Training Strategy

Visual Encoding Feature Layer Comparison Experiment

The visual encoder extracts image features and outputs a total of 24 hidden layers (disregarding the hidden layer feature values before being input into the visual encoder), and it is necessary to select the appropriate layer as the visual feature information input. Therefore, comparison experiments were conducted on 3 datasets, and the last 5 hidden layers (20th to 24th) features were selected as the visual feature information to train the TongueVLM model; the accuracies on the test datasets are shown in Table 2 below. The experimental results show that the model performs best in all tasks when the penultimate layer of features output from the visual encoder, that is, the 23rd layer of features, is selected as the final visual feature information. These findings reveal that the visual features extracted at that depth level have significant advantages for multimodal intelligent analysis related to TCM tongue diagnosis.

Table 2. Accuracy of different hidden layers as a visual encoder in the test dataset. A total of 500 samples were randomly selected from each of the 3 multimodal test datasets for experimentation. It shows the accuracy of different hidden layers as visual encoders in the test dataset, with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets).

Hidden layers	TCM-SFT[a]-Description-103k		TCM-SFT-Healthstatus-103k		TCM-SFT-Communication-103k	
	Test result details	Accuracy (%)	Test result details	Accuracy (%)	Test result details	Accuracy (%)
Layer-24	54,102,344/500	79	44,29,427/500	88.3	60,29,411/500	85.1
Layer-23	51,100,349/500	79.8	44,26,430/500	88.6	60,23,417/500	85.7
Layer-22	52,105,343/500	79.1	45,37,418/500	87.3	63,25,412/500	84.9
Layer-21	52,114,334/500	78.2	45,49,406/500	86.1	63,32,405/500	84.2
Layer-20	55,117,328/500	77.3	48,48,404/500	85.6	65,47,388/500	82.3

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

Modal Fusion Module Structure Experiment

To verify the reasonableness of the structure of the modal fusion module, 3 different neural network structures, namely, numbers 1 (Linear + GELU), 2 (Linear + GELU + Linear + GELU + Linear), and 3 (Linear + GELU + Linear + GELU + Linear + Linear + Res), are designed, and initial weights are randomized with the same datasets; the accuracy on the test datasets is shown in Table 3. The results show that the difference in the effect of the 3 structures on the final performance is not significant. However, the use of the fusion structure of group number 3, containing multilayer linear transformation, together with a multilayer nonlinear activation function and the introduction of residual connections, slightly outperforms the performance of the other schemes. Therefore, group number 3 is chosen as the neural network structure for modal fusion learning.

Table 3. Comparative analysis of the modal fusion neural network structure in terms of accuracy on the test datasets. A total of 500 samples were randomly selected from each of the 3 multimodal test datasets for experimentation with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets.

Number	TCM-SFT[a]-Description-103k		TCM-SFT-Healthstatus-103k		TCM-SFT-Communication-103k	
	Test result details	Accuracy (%)	Test result details	Accuracy (%)	Test result details	Accuracy (%)
1	55,99,346/500	79.1	46,26,428/500	88.2	62,27,411/500	84.9
2	54,98,348/500	79.4	45,29,426/500	88.1	60,27,413/500	85.3
3	51,100,349/500	79.8	44,26,430/500	88.6	60,23,417/500	85.7

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

Modal Fusion Module Pretraining Experiment

The weight parameters of the modal fusion are also fine-tuned for training during the overall training of the TongueVLM model; thus, the pretraining of the modal fusion is compared to that of number 1, the pretraining is fine-tuned individually, and number 2 is randomly initialized in TongueVLM without pretraining. The accuracy of the 3 test datasets is shown in Table 4 below. The results show that the accuracy improvement using the individually fine-tuned pretraining approach is significant, with 2.3%, 5%, and 3.5% improvement for the 3 datasets, respectively. Fine-tuning the modal fusion layer using the Chinese interpretation feature description data allows the model to focus more on constructing and optimizing higher-order correspondences between visual-language features.

Table 4. Evaluation of pretraining methods for modal fusion modules. A total of 500 samples were randomly selected from each of the 3 multimodal test datasets for experimentation with detailed data represented as: number of scores 0, number of scores 50, number of scores 100/total number of test datasets.

Numbers	TCM-SFT[a]-Description-103k		TCM-SFT-Healthstatus-103k		TCM-SFT-Communication-103k	
	Test result details	Accuracy (%)	Test result details	Accuracy (%)	Test result details	Accuracy (%)
1	51,100,349/500	79.8	44,26,430/500	88.6	60,23,417/500	85.7
2	53,119,328/500	77.5	46,73,381/500	83.5	63,52,385/500	82.2

[a]TCM-SFT: traditional Chinese medicine–supervised fine-tuning.

Analysis of Module Effectiveness and Interpretability

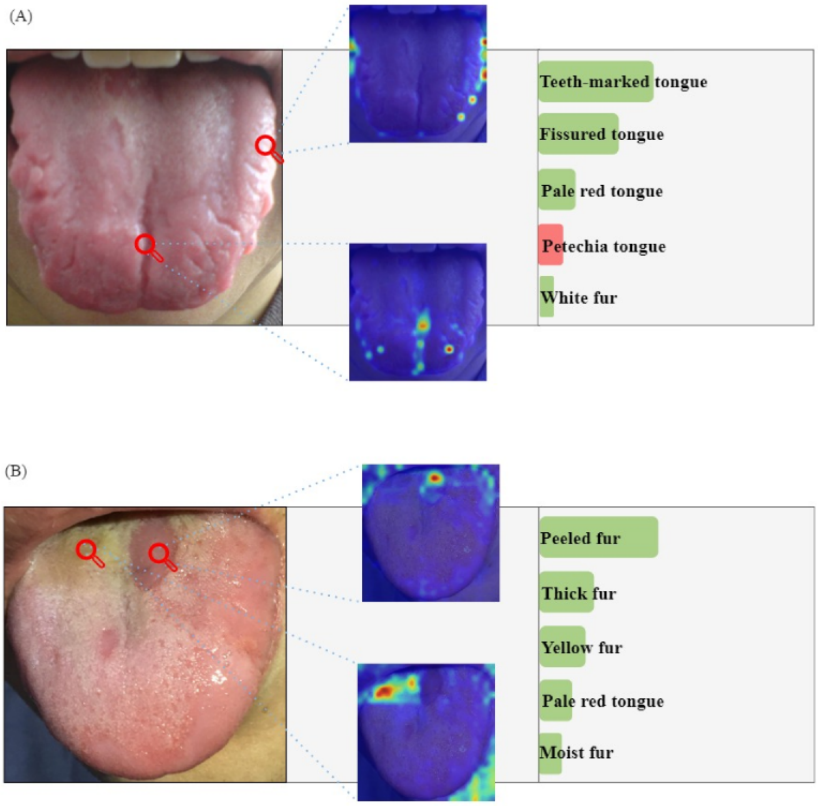
In this experiment, the visualization of the projection of the attention weight values of the multihead self-attention layer during the forward computation of the visual encoder to the image is investigated, and pseudo-color maps of the regional thermal response during the visual encoding process for typical texture feature locations on the surface of the tongue image are generated.

The multihead attention weights during the forward propagation of the cached visual encoder are $Atten\_weights \in R^{(num\_heads,dim,dim)}$, where $num\_heads \in (1,2,\ldots,16)$ is the number of heads of multihead attention, and *dim*=577 is the length of the encoding sequence. The 0th position of the encoding sequence is the classification head, and the $E_i \in R^{(dim-1,dim-1)}$ vector represents the correlation between the 24 × 24 patches of the image in the ith attention header.

As shown in Figure 5, feature locations were selected in the tongue image; Figure 5A for fissured and teeth-marked tongues, and Figure 5B for peeled fur and yellow moss. The attention weights for the specified location are projected into the pseudo-color map of the original image. The text labels of the top 5 tongue features are obtained as output, with red indicating incorrect prediction and green indicating correct prediction, and the length of the color indicates the magnitude of its probability.

Through the feature attention visualization map, it can be seen that the visual encoder for tongue texture features focuses on localized regional positional responses, while the tongue color feature focuses on the overall area response of the image, so that the visual encoder can learn the effective feature information on the tongue, which provides good support for the subsequent feature fusion.

Figure 5. Visualization of the thermal response of the visual encoder's multihead attention. Presents the top 5 categories and their confidence levels of tongue image features predicted by the model in text form. Green represents correctly predicted labels, red represents incorrectly predicted labels, and the length of the color bars intuitively indicates the probability of the prediction.



Validation of Multimodal Capabilities Across Tasks

Tongue Image Description Generation Capability

In the task of describing the tongue images, we observe that LLaVA-1.6, LLaVA-OneVision, and Qwen2.5-VL-7B all generate output answers according to structured texts. Moreover, the text of LLaVA-1.6's response to the tongue surface image was intermingled with the description of the tongue sublingual veins. This type of response may be a language hallucination of the LLM. The output text of the TongueVLM is relatively short, but in terms of content, it focuses mainly on image color, texture, and morphological information. Representative examples of model outputs are shown in Table S1 in Multimedia Appendix 1. The tongue feature interpretation of the TongueVLM model is closer to the results of visual model recognition, which fully demonstrates that the TongueVLM model, which has been fine-tuned with the tongue multimodal datasets, solves the conceptual alignment between the visual encoding of the images and the TCM terminology.

Physical Constitution Reasoning and Analysis Capability

In the task of physical constitution reasoning about the tongue image, we observe that the output of the LLaVA-1.5 model completely deviates from the TCM context and seems to lack an understanding of the concept of TCM physique, and the LLaVA-1.6 model outputs the reasoning thinking about TCM physique through the tongue image. The LLaVA-OneVision model misjudges the tongue feature as a red tongue, but can reason about the physical constitution in combination with the tongue feature. Of course, the output is also wrong. QWen2.5-VL-7B still follows some kind of templated output content, which is accurate in understanding the tongue color and color of the tongue fur features, but misclassifies the tongue shape as fat and large and dentate to the extent that the analysis of the physical constitution is not accurate enough.

In contrast, the TongueVLM model shows professional advantages in the TCM field in terms of effectively recognizing the tongue features and then briefly outputs the results through the tongue features, but its language expression level is slightly rough and needs to be further improved. Representative examples of model outputs are shown in Table S2 in Multimedia Appendix 1.

Tongue Image Dialogue Capability

In the multiround dialogue task of TCM with tongue images, we observe that all the models can accurately recognize the teeth-marked tongue feature in the tongue image in the first round of dialogue, and LLaVA-1.6 not only interprets the tongue feature dimension but also associates more scientific knowledge with the physical condition. However, in the second round of dialogue, LLaVA-1.5 and LLaVA-1.6 gave vague answers of "possibly related." Moreover, LLaVA-OneVision and Qwen2.5-VL-7B were able to provide definite answers, but they were not related to the teeth-marked tongue feature in the previous round of dialogue. The TongueVLM model combines the content of the teeth-marked tongue feature and provides the exact answer; thus, the TongueVLM model still outperforms the other models. Representative examples of model outputs are shown in Table S3 in Multimedia Appendix 1.

The experimental results show that the TongueVLM model outperforms the general-purpose large model on all 3 tasks. First, in terms of tongue feature descriptions, the TongueVLM model is closest to the recognition results of the visual model and includes color and morphological features and local positional features, and the parallel computation time of multiple visual models is greater than that of the large model, which fully reflects the image feature retention ability and computational performance of the TongueVLM visual encoder. Second, in terms of physical constitution reasoning, it is obvious that LLaVA-1.5 does not have expertise in TCM. LLaVA-1.6, LLaVA-OneVision, and Qwen2.5-VL-7B are generated by interpreting the text according to the unified paradigm; it cannot distinguish between tongue photos and sublingual photos, and identifying the positive and negative features is difficult. The TongueVLM model, on the other hand, quantitatively analyzes and localizes the positive features of the tongue image and can reason about the physical constitution based on the tongue image features, with an accuracy of 78.6%. The traditional method requires multiple visual models to extract the tongue features and then input the feature data into decision trees, support vector machines, and other classifiers for supervised learning of somatic qualities, which is less efficient to execute. Third, in terms of multiround Q&As, the TongueVLM model's interactive answers are brief, the contextual correlation is stronger, and the language discourse is not as strong as that of LLaVA-1.6, but its overall answers are more precise and concise, and language generation is not an advantage of the visual model.

Discussion

Principal Findings

This study proposes a TongueVLM network architecture design and a 3-stage training framework, achieving state-of-the-art performance on multimodal tongue image diagnosis tasks. Crucially, ablation experiments confirm that the staged training strategy is essential for stable optimization, while visual analysis demonstrates that the model spontaneously focuses on tongue regions relevant to clinical diagnosis.

The TongueVLM multimodal model proposed in this study is characterized by the following key features. First, the visual encoder characterizes the image features with the language encoding of low-dimensional vectors and approximates the feature extraction capability of the visual model by migration learning of TCM images from the CLIP-ViT pretrained model, which is a key step in the visual-to-large language model. Second, the modal fusion module splices visual and textual instruction data, realizes a seamless connection between natural language and image feature information, and unifies the multimodal data format input to the large model. Third, the TongueVLM model is fine-tuned and trained end-to-end on the TCM dataset. The model exhibits high accuracy in all 3 tasks of tongue feature description generation, physical constitution reasoning, and multiround dialogue.

Limitations and Future Work

This study has several limitations. First, the current data modality only includes tongue image and text-based question-answering data, failing to encompass the complete information system of TCM's "observation, auscultation, inquiry, and palpation." Second, the dataset's diversity requires enhancement, with shortcomings in ethnic distribution and standardization of image acquisition environments, which may compromise the model's generalizability. Third, evaluation of generated texts (eg, tongue pattern descriptions) still relies on expert judgments with inherent subjectivity, lacking objective, fine-grained automated assessment criteria. Finally, the model's generalization capability for clinically rare tongue patterns remains to be validated.

Future research may explore the following directions: first, in terms of model architecture, a gating mechanism is introduced to achieve adaptive attention regulation. Meanwhile, it is considered to incorporate modules such as depth-wise separable convolution to compensate for the shortcomings of self-attention mechanisms in capturing local detail features. Second, expand multimodal data sources by systematically integrating comprehensive information from facial diagnosis, pulse diagnosis, olfactory diagnosis, and physical constitution identification to construct a full-disease-course diagnostic model more aligned with TCM's holistic perspective. Third, fuse multidimensional data acquisition technologies such as 400-1000 nm spectral data and thermal infrared imaging data, and 3D tongue morphology to establish a more refined quantitative tongue pattern characterization system. Fourth, explore automated evaluation frameworks based on LLMs to achieve fine-grained quantitative assessments of the professionalism, consistency, and logical coherence of generated content. Fifth, advance prospective trials in collaboration with clinical institutions to validate the model's auxiliary diagnostic efficacy and clinical applicability in real-world diagnostic settings. Sixth, currently, TongueVLM is built upon a 7B-parameter model, and its deployment still imposes certain computational requirements. We will explore parameter-efficient fine-tuning techniques (eg, low-rank adaptation) and model quantization. Finally, the model will incorporate continuous learning mechanisms to iteratively improve its tongue image recognition accuracy and diagnostic reasoning ability, thereby advancing research on artificial intelligence–assisted TCM tongue diagnosis.

Conclusions

Intelligent diagnosis and treatment technology in the field of TCM is often based on knowledge graphs, deep learning visual models, and machine learning approaches. Traditional supervised learning is limited by the quality and size of the training data, as well as the difficulty of fusion with natural language and its limited generalizability. The findings demonstrate that the TongueVLM model achieves ideal results in terms of understanding tongue images, constitutive reasoning, and dialogue abilities, and provides a foundational framework suitable for future integration into a Chinese medicine intelligent diagnosis system.

Funding

We are thankful for funding support from the Project of Plans for Major Provincial Science & Technology (202303a07020008) and Xin'an Institute of Medicine and Modernization of TCM Research and Development Project: "Leading the Charge with Open Competition" (2024CXMMTCM002).

Data Availability

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to this study's conception and design. Material preparation, data collection, and analysis were performed by CP, JG, NY, YW, RC, and CD. The first draft of this paper was written by CP, and all authors commented on previous versions of this paper. All authors read and approved this final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Tables (S1-S3) and figures (S1-S4), formulas, and example data.
[DOCX File , 33112 KB-Multimedia Appendix 1]

References

1. Pak S, Son HJ, Kim D, Woo JY, Yang IK, Hwang HS, et al. Comparison of CNNs and transformer models in diagnosing bone metastases in bone scans using Grad-CAM. Clin Nucl Med. 2025;50(7):596-604. [doi: 10.1097/RLU.0000000000005898] [Medline: 40237349]
2. Zhang Y, Wang J, Górriz JM, Wang S. Deep learning and vision transformer for medical image analysis. J Imaging. 2023;9(7):147. [FREE Full text] [doi: 10.3390/jimaging9070147] [Medline: 37504824]

3. Nurgazin M, Tu NA. A comparative study of vision transformer encoders and few-shot learning for medical image classification. 2023. Presented at: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2025 Oct 19:2505-2513; Honolulu, Hawaii. [doi: 10.1109/iccvw60793.2023.00265]
4. Dayan B. Lung disease detection with vision transformers: a comparative study of machine learning methods. arXiv. Preprint posted online on Nov 18, 2024. [doi: 10.48550/arXiv.2411.11376]
5. Falqueto P, Sanfeliu A, Palopoli L, Fontanelli D. Learning priors of human motion with vision transformers. In: IEEE. 2025. Presented at: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC); July 02-04, 2024:1-9; Osaka, Japan. [doi: 10.1109/COMPSAC61105.2024.00060]
6. Zhang H, Ramachandra R, Raja K, Busch C. Generalized single-image-based morphing attack detection using deep representations from vision transformer. 2024. Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 17, 2024:1510-1518; Seattle, WA. [doi: 10.1109/cvprw63382.2024.00158]
7. Sheffield B, Ellen J, Whitmore B. On vision transformers for classification tasks in side-scan sonar imagery. arXiv. Preprint posted online on Sep 18, 2024. [doi: 10.48550/arXiv.2409.12026]
8. Gao W, Han B, Sun Z, Yan Y, Ye Y, Feng J, et al. A novel system for precise identification and explainability analysis based on multimodal learning combining laser-induced breakdown spectroscopy and laser-induced plasma acoustic signals. Talanta. 2025;293:128182. [doi: 10.1016/j.talanta.2025.128182] [Medline: 40252502]
9. Singh G. Comparative analysis of vision transformers and traditional deep learning approaches for automated pneumonia detection in chest X-Rays. arXiv. Preprint posted online on Jul 11, 2025. [doi: 10.48550/arXiv.2507.10589]
10. Alqutayfi A, Almattar W, Al-Azani S, Khan FA, Qahtani AA, Alageel S, et al. Explainable disease classification: exploring Grad-CAM analysis of CNNs and ViTs. JAIT. 2025;16(2):264-273. [doi: 10.12720/jait.16.2.264-273]
11. Zhang W, Belcheva V, Ermakova T. Interpretable deep learning for diabetic retinopathy: a comparative study of CNN, ViT, and hybrid architectures. Computers. 2025;14(5):187. [doi: 10.3390/computers14050187]
12. Amangeldi A, Taigonyrov A, Jawad MH, Mbonu CE. CNN and ViT efficiency study on tiny ImageNet and DermaMNIST datasets. arXiv. Preprint posted online on Feb 13, 2026. [doi: 10.48550/arXiv.2505.08259]
13. Mehdipour S, Mirroshandel SA, Tabatabaei SA. Vision transformers in precision agriculture: a comprehensive survey. Intell Syst Appl. 2026;29:200617. [doi: 10.1016/j.iswa.2025.200617]
14. Agarwal A, Gearon J, Rank R, Chenevert E. Fighting fires from space: leveraging vision transformers for enhanced wildfire detection and characterization. arXiv. Preprint posted online on Apr 18, 2025. [doi: 10.48550/arXiv.2504.13776]
15. Haotian L, Chunyuan L, Qingyang W, Lee YL. Visual instruction tuning. 2023. Presented at: Proceedings of the 37th International Conference on Neural Information Processing Systems; December 10-16, 2023:1516; New Orleans, LA. [doi: 10.5860/choice.189890]
16. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. 2024. Presented at: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 16-22, 2024:26286-26296; Seattle, WA. [doi: 10.1109/cvpr52733.2024.02484]
17. Li B, Zhang Y, Guo D. LLaVA-OneVision: easy visual task transfer. arXiv. Preprint posted online on Oct 26, 2024. [doi: 10.48550/arXiv.2408.03326]
18. Yang Z, Li L, Lin K. The dawn of LMMs: preliminary explorations with GPT-4V(ision). arXiv. Preprint posted online on Oct 11, 2023. [doi: 10.48550/arXiv.2309.17421]
19. Gemini T, Georgiev P, Lei V. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv. Preprint posted online on Dec 16, 2024. [doi: 10.48550/arXiv.2403.05530]
20. Dehghani M, Djolonga J, Mustafa B. Scaling vision transformers to 22 billion parameters. arXiv. Preprint posted online on Feb 10, 2023. [doi: 10.48550/arXiv.2302.05442]
21. Xuyan H, Meng S, Chengxing S, Haoxuan L, Jianlin Z. Visual-language reasoning large language models for primary care: advancing clinical decision support through multimodal AI. Vis Comput. 2025;41(13):11327-11348. [doi: 10.1007/s00371-025-04109-y]
22. Thapa S, Naseem U, Zhou L. Vision-language models for biomedical applications. 2024. Presented at: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications; Oct 28-Nov 1, 2024:1-2; Melbourne, Victoria, Australia. [doi: 10.1145/3689096.3690770]
23. Raminedi S, Shridevi S, Won D. Multi-modal transformer architecture for medical image analysis and automated report generation. Sci Rep. 2024;14(1):19281. [FREE Full text] [doi: 10.1038/s41598-024-69981-5] [Medline: 39164302]
24. AlShibli A, Bazi Y, Rahhal MMA, Zuair M. Vision-BioLLM: large vision language model for visual dialogue in biomedical imagery. Biomed Signal Proc Control. 2025;103:107437. [doi: 10.1016/j.bspc.2024.107437]
25. Kalpelbe BC, Adaambiik AG, Peng W. Vision language models in medicine. arXiv. Preprint posted online on Feb 24, 2025. [doi: 10.48550/arXiv.2503.01863]
26. Peng C, Zhang K, Lyu M, Liu H, Sun L, Wu Y. Scaling up biomedical vision-language models: fine-tuning, instruction tuning, and multi-modal learning. J Biomed Inform. 2025;171:104946. [doi: 10.1016/j.jbi.2025.104946] [Medline: 41138953]
27. Liu Y, Fan L, Zhao M, Wei D, Zhao M, Dong Y, et al. Study on a traditional chinese medicine constitution recognition model using tongue image characteristics and deep learning: a prospective dual-center investigation. Chin Med. 2025;20(1):84. [FREE Full text] [doi: 10.1186/s13020-025-01126-w] [Medline: 40506765]

28. Jiatuo XU, Tao J, Shi L. Research status and prospect of tongue image diagnosis analysis based on machine learning. Digital Chin Med. 2024;7(1):3-12. [doi: 10.1016/j.dcmed.2024.04.002]
29. Tian-Yu GU, Zhuang-Zhi Y, Jie-Hui J. Classifying Chinese medicine constitution using multimodal deep-learning model. Chin J Integr Med. 2024;30(2):163-170. [doi: 10.1007/s11655-022-3541-8]
30. Ye Q, Liu J, Chong D. Qilin-Med: Multi-stage knowledge injection advanced medical large language model. arXiv. Preprint posted online on Apr 17, 2024. [doi: 10.48550/arXiv.2310.09089]
31. Jia Y, Ji X, Wang X, Zhang H, Meng Z, Zhang J, et al. Qibo: a large language model for traditional chinese medicine. Expert Syst Appl. 2025;284:127672. [doi: 10.1016/j.eswa.2025.127672]
32. Yu S, Xu X, Xu F. Enhancing the traditional chinese medicine capabilities of large language model through reinforcement learning from AI Feedback. arXiv. Preprint posted online on Nov 1, 2024. [doi: 10.48550/arXiv.2411.00897]
33. Xie J, Yu Y, Zhang Z. TCM-Ladder: a benchmark for multimodal question answering on traditional chinese medicine. arXiv. Preprint posted online on Oct 24, 2025. [doi: 10.48550/arXiv.2505.24063]
34. Sun Z, Huang R, Feng J. DoPI: doctor-like proactive interrogation LLM for traditional chinese medicine. arXiv. Preprint posted online on Jul 7, 2025. [doi: 10.48550/arXiv.2507.04877]
35. Wei S, Peng X, Wang Y, Shen T, Si J, Zhang W, et al. BianCang: a traditional chinese medicine large language model. IEEE J Biomed Health Inf. 2025:1-12. [doi: 10.1109/jbhi.2025.3612415]
36. Khattak MU, Kunhimon S, Naseer M. UniMed-CLIP: towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv. Preprint posted online on Dec 13, 2024. [doi: 10.48550/arXiv.2412.10372]
37. LASA Team, Xu W, Chan HP, Li L. Lingshu: a generalist foundation model for unified multimodal medical understanding and reasoning. arXiv. Preprint posted online on Jun 13, 2025. [doi: 10.48550/arXiv.2506.07044]
38. Alsinglawi B, Mccarthy C, Webb S, Fluke C, Saidy NT. A lightweight large vision-language model for multimodal medical images. arXiv. Preprint posted online on Apr 8, 2025. [doi: 10.48550/arXiv.2504.05575]
39. Wang Y, Liu J, Gao S, Feng B, Tang Z. V2T-CoT: from vision to text chain-of-thought for medical reasoning and diagnosis. Med Image Comput Comput Assisted Intervention – MICCAI 2025. 2026;15964:658-668. [doi: 10.1007/978-3-032-04971-1_62]
40. Li C, Wong C, Zhang S. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. arXiv. Preprint posted online on Jun 1, 2023. [doi: 10.48550/arXiv.2306.00890]
41. Yang X, Miao J, Yuan Y, Jiaze W, Qi D. Medical large vision language models with multi-image visual ability. Med Image Comput Comput Assisted Intervention – MICCAI 2025. 2026;15964:402-412. [doi: 10.1007/978-3-032-04971-1_38]
42. Xing Q, Song Z, Zhang Y. MCA-RG: enhancing LLMs with medical concept alignment for radiology report generation. Med Image Comput Comput Assisted Intervention – MICCAI 2025. 2026;15964:380-390. [doi: 10.1007/978-3-032-04971-1_36]
43. Yang Y, Ma T, Li R, Zheng X, Shan G. JingFang: an expert-level large language model for traditional chinese medicine clinical consultation and syndrome differentiation-based treatment. arXiv. Preprint posted online on Jan 20, 2026. [doi: 10.48550/arXiv.2502.04345]
44. Xu Z, Li Q, Nie W, Wang W, Liu A. Structure causal models and LLMs integration in medical visual question answering. IEEE Trans Med Imaging. 2025;44(8):3476-3489. [doi: 10.1109/tmi.2025.3564320] [Medline: 40299735]
45. Traditional Chinese Medicine Identification v2 API. AI Open Platform for TCM Tongue Diagnosis. Beijing. Hefei Cloud Diagnostics Information Technology Co; 2025. URL: https://www.ai-tongue.com/doc_n4t9r9w8/apiapp/composite/home.html [accessed 2026-02-24]

Abbreviations

API: application programming interface
CLIP-ViT: Contrastive Language-Image Pre-Training With Vision Transformer
GELU: Gaussian Error Linear Unit
LLaMA: large language model meta artificial intelligence
LLM: large language model
Q&A: question and answer
TCM: traditional Chinese medicine
TCM-SFT: traditional Chinese medicine–supervised fine-tuning
VLM: visual language model

Edited by A Benis; submitted 06.Nov.2025; peer-reviewed by Y Liu, M Fang; comments to author 05.Dec.2025; revised version received 17.Dec.2025; accepted 17.Feb.2026; published 12.Mar.2026

Please cite as:
Peng C, Gao J, Yang N, Wang Y, Chen R, Dong C
Application of a Large Visual Language Model on Tongue Image Description Generation and Physical Constitution Reasoning in Traditional Chinese Medicine (TongueVLM): Model Development and Validation Study
JMIR Med Inform 2026;14:e87237
URL: https://medinform.jmir.org/2026/1/e87237
doi: 10.2196/87237
PMID: