Scientific Research Publishing

Health, 2026, 18(11), 1012-1023
https://www.scirp.org/journal/health
ISSN Online: 1949-5005
ISSN Print: 1949-4998

Evaluation of Vision-Language Models for Traditional Chinese Medicine Tongue Diagnosis: A Comparative Study of AI Performance against Human Expert Assessment

Ryan Kern, Brian Z. Ring, Huijun Z. Ring*

Stanford University School of Medicine, Stanford, CA, USA
Email: ryankern@stanford.edu, bzring@stanford.edu, *ring@stanford.edu

How to cite this paper: Kern, R., Ring, B.Z. and Ring, H.Z. (2026) Evaluation of Vision-Language Models for Traditional Chinese Medicine Tongue Diagnosis: A Comparative Study of AI Performance against Human Expert Assessment. *Health*, **18**, 1012-1023.
https://doi.org/10.4236/health.2026.1811062

Received: October 26, 2026
Accepted: November 14, 2026
Published: November 17, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Background: Traditional Chinese Medicine (TCM) tongue diagnosis is a complex diagnostic practice requiring years of training. Recent advances in vision-language models (VLMs) offer potential for automated assistance, though their capability for nuanced TCM assessment remains unexplored. This preliminary study investigates whether modern VLMs can perform TCM tongue feature assessment with agreement approaching that observed between human evaluators. **Methods:** A small dataset of standardized tongue images, contributed voluntarily with informed consent, was independently scored by two human evaluators and three VLMs using a consistent, structured prompt. Nine key diagnostic features were evaluated on a 0 - 100 scale across four tongue regions (base, center, sides, tip): tongue body color, shape, swelling, coating color, coating thickness, coating distribution, cracks or fissures, spots, and tooth marks. Agreement was measured using mean absolute deviation (MAD), root mean squared error (RMSE), bias, and categorical agreement of binned scores. **Results:** Inter-rater agreement between human evaluators established a benchmark (MAD: 14.4, agreement of binned scores: 66.5%). All three VLMs demonstrated measurable capability for TCM tongue assessment, with varying performance levels. The best-performing model achieved agreement approaching the human inter-rater benchmark (MAD: 18.0, agreement: 62.8%), while other models showed MAD values ranging from 18.8 to 37.9. A single standardized prompt was used across all models without model-specific optimization, which likely limited individual model performance. Performance patterns remained consistent when restricted to features with positive Lin

CCC between human evaluators, suggesting similar performance across features of differing reliability. **Conclusions:** This small preliminary study demonstrates that vision-language models possess fundamental capabilities for Traditional Chinese Medicine tongue diagnosis, with agreement approaching that observed between human evaluators. The observed differences between models should be interpreted cautiously, as model-specific prompt optimization was not performed and could substantially alter relative performance. These findings establish proof-of-concept that VLMs can perform complex, nuanced diagnostic assessments requiring cultural medical knowledge and subtle visual discrimination, suggesting promising applications for clinical decision support, education, and diagnostic standardization in TCM practice.

Keywords

Traditional Chinese Medicine, Tongue Diagnosis, Vision-Language Models, Artificial Intelligence, Medical Image Analysis, Gemini, Qwen-VL, Llava

1. Introduction

1.1. Traditional Chinese Medicine Tongue Diagnosis

Traditional Chinese Medicine (TCM) tongue diagnosis represents a sophisticated diagnostic modality with over two millennia of documented practice. Practitioners assess subtle variations in tongue body color, coating characteristics, shape, texture, and other features to inform diagnostic and therapeutic decisions. The tongue is considered a mirror of internal organ function, with specific tongue regions corresponding to different organ systems according to TCM theory [1]. Despite its long history and widespread clinical use, tongue diagnosis remains a largely subjective practice requiring extensive training and experience, leading to inter-practitioner variability and challenges in standardization.

Contemporary efforts to objectify and standardize tongue diagnosis have employed various imaging and computational approaches [2]. Early studies focused on color space quantification and texture analysis using traditional computer vision techniques. More recently, deep learning methods have shown promise for automated tongue feature extraction and classification. However, these approaches typically require large labeled datasets for specific classification tasks and may not capture the nuanced, multi-dimensional nature of expert TCM assessment.

1.2. Previous Artificial Intelligence Approaches to Tongue Diagnosis

Artificial intelligence applications in tongue diagnosis have evolved considerably over the past decade. Convolutional neural networks (CNNs) have demonstrated substantial success in specific classification tasks such as tongue coating classification and tongue body color categorization [3]-[5].

However, these CNN-based approaches face several limitations. First, they require extensive task-specific training data, which is resource-intensive to collect and annotate. Second, they typically address single classification tasks (e.g., coating thickness) rather than comprehensive multi-feature assessment. Third, they cannot easily incorporate the contextual reasoning and holistic pattern recognition that characterizes expert TCM practice. Finally, adapting these models to new features or assessment criteria requires complete retraining with new labeled datasets.

1.3. Vision-Language Models: A New Paradigm

The emergence of large-scale vision-language models (VLMs) represents a paradigm shift in medical image analysis. These models, trained on massive datasets of images paired with natural language descriptions, can perform complex visual reasoning tasks through natural language prompting without task-specific training. Recent VLMs such as Google's Gemini, Alibaba's Qwen-VL, and the open-source Llava have demonstrated impressive capabilities across diverse visual understanding tasks [6]-[8].

In medical imaging contexts, VLMs offer several potential advantages. They can perform zero-shot analysis of novel imaging modalities through appropriate prompting. They can simultaneously assess multiple features in a single analysis, mirroring clinical practice. They can incorporate domain knowledge and reasoning patterns described in natural language. Perhaps most importantly, they can provide detailed explanations of their assessments, enabling transparency and educational value.

Recent work has begun exploring VLM applications in medical imaging, including radiology image interpretation, dermatological diagnosis and even in TCM applications [9]. However, few studies have systematically evaluated VLM performance for Traditional Chinese Medicine tongue diagnosis—a particularly challenging domain requiring cultural medical knowledge, subtle visual discrimination, and integration of multiple feature assessments.

1.4. Study Objectives

This study aims to investigate whether state-of-the-art vision-language models can score visual Traditional Chinese Medicine tongue features in agreement with a human evaluator, and how this agreement compares with that observed between two human evaluators. We evaluate three prominent VLMs—Google's Gemini, Alibaba's Qwen-VL, and the open-source Llava—using a standardized dataset and structured prompting approach. Our specific objectives are:

- To establish baseline inter-rater agreement between experienced human TCM practitioners
- To evaluate VLM performance across multiple tongue diagnostic features using quantitative agreement metrics
- To compare VLM performance on easily-measured versus difficult-to-assess

features

- To assess whether observed performance differences between models reflect true capability differences or limitations of the standardized prompting approach

We hypothesize that modern VLMs possess sufficient visual understanding and reasoning capabilities to perform multi-feature tongue assessment with agreement approaching that observed between human evaluators, though performance may vary across models depending on their training and architectural characteristics.

2. Materials and Methods

2.1. Study Design

This was a cross-sectional diagnostic agreement study comparing vision-language model performance against human expert assessment for Traditional Chinese Medicine tongue diagnosis. The dataset comprised five standardized tongue images. The images were provided voluntarily with consent for their use in this study. Images were captured using consistent protocols: participants were asked to extend their tongue naturally without excessive tension, under standardized lighting conditions (daylight-balanced LED, 5000 K color temperature), with the camera positioned approximately 20 cm from the tongue at a perpendicular angle.

2.2. Evaluated Features and Tongue Regions

Nine key TCM tongue diagnostic features were evaluated in this study across four distinct tongue regions. The assessed features were:

- Tongue body color: Assessment of the underlying tongue tissue color abnormality
- Tongue shape: Analysis of overall tongue morphology and size
- Swelling: Evaluation of tongue enlargement or puffiness
- Coating color: Evaluation of the tongue coating appearance
- Coating thickness: Measurement of tongue coating density and opacity
- Coating distribution: Assessment of coating uniformity and location patterns
- Cracks or fissures: Identification and assessment of tongue surface fissures
- Spots: Detection of discolored spots or petechiae on the tongue surface
- Tooth marks: Assessment of indentations along tongue edges from dental impressions

Each feature was evaluated independently across four tongue regions corresponding to traditional TCM organ system mapping:

- Base (root): Posterior portion, associated with kidney, bladder, and intestines
- Center: Middle portion, associated with spleen and stomach
- Sides (edges): Lateral portions, associated with liver and gallbladder
- Tip: Anterior portion, associated with heart and lungs

This region-based assessment approach aligns with traditional TCM diagnostic theory, which associates specific tongue areas with different organ systems and

physiological functions. Each feature was scored on a continuous 0 - 100 scale for each region, allowing for both regional specificity and overall tongue assessment through aggregation of regional scores.

2.3. Human Expert Evaluation

Two human evaluators independently assessed all tongue images. The primary evaluator (Evaluator 1) served as the comparison reference against which all other assessments were compared. The secondary evaluator (Evaluator 2) provided an independent assessment to establish inter-rater reliability benchmarks.

Evaluators scored each of the nine features on a 0 - 100 continuous scale for each of the four tongue regions. They were provided with detailed scoring guidelines including photographic reference examples, but were instructed to use their clinical judgment in assigning scores. Evaluators were blinded to each other's scores and to the AI model outputs. Scoring was conducted over a two-week period, with images presented in randomized order to minimize fatigue and order effects.

2.4. Vision-Language Model Evaluation

Three vision-language models were evaluated: Google Gemini (version: gemini-pro-vision, accessed November 2024), Alibaba Qwen-VL (version: Qwen-VL-Chat, accessed November 2024), and Llava (version: llava-v1.5-13b, accessed November 2024). These models were selected to represent different development approaches: commercial closed-source (Gemini), commercial open-weights (Qwen-VL), and fully open-source (Llava).

A critical design decision was the use of a single, standardized prompt across all three models to ensure fair comparison. This prompt was structured to provide clear instructions about the nine features, the four regions, the 0 - 100 scoring scale, and guidance for visual assessment. However, this standardized approach represents an important limitation: the prompt was not optimized for each model's specific strengths and characteristics. Model-specific prompt engineering could potentially improve individual model performance substantially. Therefore, the observed performance differences should be interpreted as reflecting capabilities under standardized prompting conditions rather than indicating absolute performance limits.

Each model received identical input: the tongue image and the standardized prompt requesting assessment of all nine features across all four regions. Model outputs were parsed to extract numerical scores. In cases where models provided qualitative descriptions rather than numerical scores, a secondary large language model (GPT-4) was used to convert descriptions to the 0 - 100 scale using consistent conversion criteria. All model evaluations were completed within a one-week period in November 2024.

2.5. Statistical Analysis

Each scored feature in each tongue region was assessed independently for each tongue image and scoring method, as well as a mean score across all tongue re-

gions. For calculating percentage agreement, scores were grouped into low, medium, and high categories using percentile bins, with less than or equal to the 50th percentile equaling low, and greater than the 50th percentile and less than the 90th percentile equaling medium, and greater than or equal to the 90th percentile equaling high.

Lin's concordance correlation coefficient (CCC) was used to compare the independent (human) evaluator to the primary (human) evaluator. Features that allowed reliable scoring between human evaluators (positive Lin CCC) were identified and used for additional comparisons of visual models to the primary evaluator.

The following metrics were calculated for each scoring method compared to the primary evaluator:

- Mean Absolute Deviation (MAD): Average of absolute differences between paired scores
- Root Mean Squared Error (RMSE): Square root of the mean of squared differences, providing greater weight to larger errors
- Bias: Mean difference between scoring methods and primary evaluator (negative values indicate systematic underestimation)
- Agreement of binned scores: Percentage of assessments falling into the same categorical bin (low/medium/high)

These metrics were calculated using all tongue features and regions, as well as in comparisons limited to those features with a positive Lin CCC between the human evaluators. All statistical analyses were performed using R version 4.3.1.

3. Results

3.1. Human Expert Agreement

The two human evaluators demonstrated substantial agreement across all tongue features, establishing a benchmark for expected performance. The independent human evaluator showed a mean absolute deviation of 14.4 from the primary evaluator, with 66.5% agreement when scores were categorized into low, medium, and high bins. The RMSE was 19.1, and the bias was −7.1, indicating a slight tendency for the independent evaluator to score features lower than the primary evaluator. When analysis was restricted to features with positive Lin's CCC between evaluators (indicating more reliable measurement), agreement improved to 69.7%, with MAD of 15.4, RMSE of 19.9, and bias of −6.8.

3.2. Vision-Language Model Performance

Table 1 presents comprehensive performance statistics for each AI model compared against the primary human evaluator. All three vision-language models demonstrated measurable capability for TCM tongue assessment, though with varying levels of agreement. One model achieved performance closely approximating human inter-rater agreement (MAD: 18.0, categorical agreement: 62.8%), while the other models showed MAD values of 18.8 and 37.9 respectively.

Table 1. Comparison of scoring methods.

Scoring Method	Tongue Features	Agreement of Binned Scores (%)	MAD	RMSE	Bias
Gemini	All features	62.8	18.0	22.6	−12.0
	Lin CCC > 0	62.7	18.0	24.3	−7.6
Llava	All features	42.6	37.9	43.8	23.4
	Lin CCC > 0	44.5	38.3	44.5	20.5
Qwen	All features	58.9	18.8	24.1	−9.8
	Lin CCC > 0	57.6	18.5	25.0	−9.0
Independent Human Evaluator	All features	66.5	14.4	19.1	−7.1
	Lin CCC > 0	69.7	15.4	19.9	−6.8

Models and an independent (human) evaluator were compared to the primary (human) evaluator. Averages of percent agreement of binned scores, MAD (Mean Absolute Deviation), RMSE (Root Mean Squared Error), and Bias (Mean difference of scoring methods to primary evaluator) per each method are shown. Results are presented for all features combined and for features with positive Lin CCC between human evaluators (Lin CCC > 0).

The observed performance differences between models are notable but should be interpreted cautiously. All models received identical standardized prompts without model-specific optimization. This design choice, while ensuring fair comparison, may have systematically advantaged or disadvantaged particular models based on their training approaches and architectural features. Model-specific prompt engineering could potentially alter the relative performance substantially, as discussed further in the limitations section.

An important pattern emerged in the bias measurements: two models showed negative bias (−12.0 and −9.8), indicating systematic underestimation of feature severity, while one model exhibited positive bias (23.4), indicating systematic overestimation. These different error patterns may reflect fundamental differences in model training or architecture, or may result from how each model interprets the standardized prompt language.

3.3. Performance on Easily-Measured versus Difficult Features

When restricting analysis to features with positive Lin's CCC between human evaluators (indicating more reliable measurement), performance patterns remained remarkably consistent across all models. For the independent human evaluator, agreement improved modestly from 66.5% to 69.7%, with MAD increasing slightly from 14.4 to 15.4. This modest improvement suggests that even "easily-measured" features in TCM tongue diagnosis maintain substantial inherent variability.

All three VLMs showed minimal performance variation between all features

and easily-measured features. The best-performing model's metrics were nearly identical (all features: MAD 18.0, agreement 62.8%; easy features: MAD 18.0, agreement 62.7%). Similarly, the other models maintained consistent performance (model 2 all features: MAD 18.8; easy features: MAD 18.5; model 3 all features: MAD 37.9; easy features: MAD 38.3).

This consistent performance across feature difficulty levels is noteworthy. It indicates that the VLMs possess robust visual understanding capabilities that generalize well across the full spectrum of diagnostic complexity. Unlike approaches that might excel on straightforward cases while failing on nuanced assessments, these models maintain stable performance regardless of measurement complexity.

4. Discussion

This study demonstrates that vision-language models possess fundamental capabilities for Traditional Chinese Medicine tongue diagnosis, with some models achieving agreement approaching that observed between human evaluators. These findings represent important proof-of-concept that VLMs can perform complex, culturally-specific medical assessments requiring subtle visual discrimination and integration of multiple diagnostic features. However, the observed performance differences between models should be interpreted cautiously given the standardized prompting approach, which may not have optimized individual model capabilities.

4.1. Comparison with Previous AI Approaches

Previous convolutional neural network approaches to tongue diagnosis have demonstrated substantial success in specific classification tasks. While these accuracy figures appear higher than the VLM agreement levels observed in this study (42% - 63% categorical agreement), direct comparison is problematic due to fundamental methodological differences. CNN studies typically evaluate binary or multi-class classification (e.g., thick vs. thin coating), whereas this study evaluated continuous 0 - 100 scale assessment across nine features and four regions simultaneously—a substantially more complex and clinically realistic task.

4.2. Interpreting Model Performance Differences

The observed performance variation between the three evaluated models ranged from near-human agreement to substantially lower agreement. However, a critical limitation must be emphasized: all models received identical standardized prompts without model-specific optimization. This design choice, while ensuring methodological fairness, likely prevented each model from achieving its optimal performance.

Different VLM architectures may respond differently to prompt structure, language style, reasoning approaches, and example formats. For instance, some models may benefit from step-by-step reasoning prompts, while others might perform better with direct instruction. Models trained on multilingual data might show superior performance when prompted in their native training languages. Models

with strong few-shot learning capabilities might benefit from example-based prompting strategies. Therefore, the performance ranking observed in this study should not be interpreted as indicating absolute superiority of certain models, but rather as reflecting capabilities under one specific prompting condition.

The primary finding is not that certain models outperform others, but rather that modern VLMs as a class possess the fundamental visual understanding and reasoning capabilities required for TCM tongue diagnosis. Future research employing model-specific prompt optimization could potentially reveal different performance relationships and might substantially improve all models' results.

4.3. Clinical Implications and Applications

The demonstration that VLMs can approach human inter-rater agreement in TCM tongue-feature scoring suggests several potential applications, pending further validation. First, these systems could serve as decision support tools for practitioners, providing second opinions or highlighting features that might be overlooked. Second, VLMs could facilitate practitioner education and training, offering consistent feedback to students learning tongue diagnosis. Third, they could contribute to diagnostic standardization, helping to reduce inter-practitioner variability in assessment.

Perhaps most importantly, VLM-based assessment tools could improve global accessibility of TCM diagnostic expertise. In regions with limited access to experienced TCM practitioners, AI-assisted tongue-feature assessment might eventually support preliminary assessment, once validated against clinical reference standards. The zero-shot nature of VLM deployment—requiring no training data or model retraining—makes this approach particularly attractive for resource-limited settings.

However, important caveats apply to clinical deployment. VLMs should be viewed as decision support tools rather than autonomous diagnostic systems. Human expert oversight remains essential, particularly for treatment decisions. Furthermore, as demonstrated by the performance variability observed in this study, careful validation and optimization of prompting strategies would be crucial before clinical deployment of any specific model.

4.4. Limitations

4.4.1. Single Standardized Prompt Approach

The most significant limitation of this study is the use of a single standardized prompt across all models without model-specific optimization. While this ensured fair comparison, it likely prevented individual models from achieving their best possible performance. Model-specific prompt engineering—leveraging each model's unique architectural features, training approaches, and reasoning capabilities—could potentially improve results substantially and might alter the relative performance rankings observed in this study. Therefore, this work should be interpreted as proof-of-concept that VLMs can perform TCM tongue diagnosis,

rather than as definitive evidence of one model's superiority over others.

4.4.2. Color Balance and Image Standardization

The analysis did not attempt to correct for differing white balance, color tints, or brightness levels between cameras using standardized aids such as color calibration cards. Color standardization is important in tongue diagnosis research, as lighting conditions and camera settings can affect perceived color. The use of color cards placed adjacent to the tongue during capture would enhance future studies, enabling more precise color correction across different imaging conditions.

At this initial stage, the primary goal was to identify which vision-language models could be effectively employed for TCM tongue diagnosis and to determine whether meaningful correlation existed between their outputs and expert human assessments. Therefore, no systematic exploration of optimal white balance or color correction methods was performed. However, the structured prompt language provided to the models did include guidance on using multiple visual features within the oral cavity to contextualize color assessment.

Notably, despite the absence of formal color correction beyond prompt-based guidance, the models demonstrated reasonable concordance with human evaluators on color-related features, suggesting some inherent robustness to lighting variations. However, future studies should systematically evaluate whether color card-based standardization would further improve model accuracy for color-dependent features.

4.4.3. Sample Size and Generalizability

The sample size (n = 5 images) was small, and the results should be regarded as preliminary. Larger multicenter validation studies with diverse patient populations, varied imaging equipment, and broader geographic representation would strengthen generalizability and enable more detailed subgroup analyses.

4.4.4. Scope of Diagnostic Assessment

This study focused exclusively on static image analysis of tongue features. Real-world TCM practice involves additional diagnostic modalities including pulse diagnosis, patient history, symptom patterns, and dynamic tongue movement observation. Future AI systems should integrate these complementary information sources for comprehensive diagnostic support. Furthermore, we did not evaluate whether AI assessments translate to appropriate treatment recommendations—a crucial step for clinical validation that requires correlation of image-based assessments with patient outcomes.

5. Conclusions

This study establishes proof-of-concept that modern vision-language models can perform Traditional Chinese Medicine tongue diagnosis with agreement approaching that observed between human evaluators. Though TCM tongue diagnosis is not an easy task—requiring subtle visual discrimination, cultural medical knowledge,

and integration of multiple features across distinct tongue regions—VLMs demonstrated measurable capability for this complex assessment.

Importantly, the observed differences between models should not be interpreted as indicating absolute superiority of certain architectures. This study employed a single standardized prompt across all models to ensure methodological fairness, but this approach likely did not optimize performance for each model's unique characteristics. Model-specific prompt engineering could potentially alter relative performance substantially. Therefore, the key finding is not that one particular model excels, but rather that VLMs as a class possess fundamental capabilities for performing nuanced TCM diagnostic assessments.

These findings have important implications for clinical practice, education, and global accessibility of TCM. VLM-based systems could serve as decision support tools for practitioners, educational aids for students, and, after clinical validation, assistance in regions with limited access to TCM expertise. The zero-shot nature of VLM deployment—requiring no training data specific to tongue diagnosis—makes this approach particularly attractive for rapid implementation and adaptation to evolving diagnostic criteria.

Future research should prioritize model-specific prompt optimization to determine the performance ceiling for each VLM architecture. Additional priorities include validation with larger, more diverse datasets; integration of color standardization protocols; incorporation of multiple diagnostic modalities beyond static images; and, critically, correlation of AI assessments with treatment outcomes and patient benefit. With these advances, VLM-based systems could become valuable tools for enhancing the precision, consistency, and accessibility of Traditional Chinese Medicine practice.

Acknowledgements

The authors thank Mofan Zhang and John Ping for their assistance with vision-language models.

Author Contributions

Ryan Kern: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, and writing—original draft preparation. Brian Z. Ring: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, and writing—original draft preparation. Huijun Z. Ring: conceptualization, methodology, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, supervision, and project administration. All authors have read and agreed to the published version of the manuscript.

AI Assisted Tool Disclosure

GPT-5.6 was used for language polishing and grammatical correction of this man-

uscript. The authors reviewed and verified all AI-generated output and assume full responsibility for the final content of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wang, Y.H. and Xu, A.L. (2014) Zheng: A Systems Biology Approach to Diagnosis and Treatments. *Science*, **346**, S13-S15.

[2] Matos, L.C., Machado, J.P., Monteiro, F.J. and Greten, H.J. (2021) Can Traditional Chinese Medicine Diagnosis Be Parameterized and Standardized? A Narrative Review. *Healthcare*, **9**, Article No. 177. https://doi.org/10.3390/healthcare9020177

[3] Wang, X., Liu, J., Wu, C., Liu, J., Li, Q., Chen, Y., *et al*. (2020) Artificial Intelligence in Tongue Diagnosis: Using Deep Convolutional Neural Network for Recognizing Unhealthy Tongue with Tooth-Mark. *Computational and Structural Biotechnology Journal*, **18**, 973-980. https://doi.org/10.1016/j.csbj.2020.04.002

[4] Li, J., Zhang, Z., Zhu, X., Zhao, Y., Ma, Y., Zang, J., *et al*. (2022) Automatic Classification Framework of Tongue Feature Based on Convolutional Neural Networks. *Micromachines*, **13**, Article No. 501. https://doi.org/10.3390/mi13040501

[5] Jiang, T., Hu, X., Yao, X., Tu, L., Huang, J., Ma, X., *et al*. (2021) Tongue Image Quality Assessment Based on a Deep Convolutional Neural Network. *BMC Medical Informatics and Decision Making*, **21**, Article No. 147. https://doi.org/10.1186/s12911-021-01508-8

[6] Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., *et al*. (2025) A Vision-Language Foundation Model for Precision Oncology. *Nature*, **638**, 769-778. https://doi.org/10.1038/s41586-024-08378-w

[7] Lin, P., Jiang, Z., Liu, Y., Li, W., Qi, Y., Sun, J., *et al*. (2025) A Vision-Language Foundation Model for Alzheimer's Disease Diagnosis Using MRI and Clinical Data. *Alzheimer's & Dementia*, **21**, e71029. https://doi.org/10.1002/alz.71029

[8] Lim, G., Elangovan, K. and Jin, L. (2024) Vision Language Models in Ophthalmology. *Current Opinion in Ophthalmology*, **35**, 487-493. https://doi.org/10.1097/icu.0000000000001089

[9] Peng, C., Gao, J., Yang, N., Wang, Y., Chen, R. and Dong, C. (2026) 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 Medical Informatics*, **14**, e87237. https://doi.org/10.2196/87237