Concept Paper

Not peer-reviewed version

Skin-TAIDE: Development of TAIDE Multimodal Models using Retrieval-Augmented Generation and Fine-Tuning Approaches for Generating Traditional Chinese Diagnosis Description of Skin Lesion

Ming-Hseng Tseng * and Jing-Wen Wu

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1897.v1

Keywords: MLLM; diagnostic description generation; Visual RAG; VLLM; TAIDE

 doi:10.20944/preprints202604.1897.v1

Concept Paper

Skin-TAIDE: Development of TAIDE Multimodal Models Using Retrieval-Augmented Generation and Fine-Tuning Approaches for Generating Traditional Chinese Diagnosis Description of Skin Lesion

Jing-Wen Wu [1] and Ming-Hseng Tseng [1,2,*]

[1] Department of Medical Informatics, Chung Shan Medical University, Taichung 402, Taiwan, R.O.C.
[2] Information Technology Office, Chung Shan Medical University Hospital, Taichung 402, Taiwan, R.O.C.
* Correspondence: mht@csmu.edu.tw; Tel.: 886-4-24730022-12214

Abstract

Purpose: With the growing interest in multimodal large language models (MLLMs) for medical image analysis, expanding the application scope of the unimodal TAIDE large-scale language model has emerged as a prominent and significant research direction. Methods: This study employed the SkinCAP multimodal dataset, which consists 4,000 images of skin lesions along with their associated textual descriptions. Two approaches for model training and evaluation are proposed: (1) A visual retrieval-augmented generation (RAG) method, which leverages transfer learning for image feature extraction and cosine similarity for image retrieval. Retrieved results are used to generate prompts that guide the TAIDE model to produce diagnostic descriptions in traditional Chinese. (2) A fine-tuning-based method that integrates the MiniGPT-V2 framework with the TAIDE model to develop a multimodal system capable of automatically generating diagnostic descriptions. Results: Model performance was evaluated using BLEU, ROUGE-L, METEOR, CIDEr, and SPICE metrics. The results demonstrate that the fine-tuning-based approach—integrating MiniGPT-V2 with the TAIDE model—achieves superior performance compared to the visual RAG-based method, which combines transfer learning-based retrieval with the TAIDE model for description generation. Conclusion: This study presents an empirical comparison of two methodologies for extending unimodal large language models into multimodal applications for the automatic generation of diagnostic descriptions of skin lesions. The findings provide valuable technical insights and serve as a reference for the development of future AI-based medical systems.

Keywords: MLLM; diagnostic description generation; Visual RAG; VLLM; TAIDE

1. Introduction

1.1. Purpose

With the rapid advancement of artificial intelligence (AI) technologies, the application of deep learning in the medical domain has attracted increasing attention. The diagnosis of skin lesions traditionally relies on the experience and professional expertise of physicians. However, due to the growing diversity of skin conditions and the increasing demand for medical services, leveraging AI to support diagnostic processes has become a critical area of research.

Conventional image classification and language generation models have been developed independently, which limits their effectiveness in multimodal tasks. As a result, multimodal models that integrate both visual and textual information have emerged as a key research focus, as they offer the potential to enhance both the efficiency and accuracy of medical diagnoses.

 doi:10.20944/preprints202604.1897.v1

This study aims to investigate the applicability of multimodal visual-language models in the context of traditional Chinese. Taking the diagnosis of skin lesions as a case study, we seek to expand the capabilities of the unimodal TAIDE language model. Furthermore, we propose the development of a Web Chat and LINE Bot service system that can process both images and text using a multimodal language model, with the ultimate goal of facilitating the practical deployment of AI in healthcare settings.

1.2. Related Work

1.2.1. TAIDE Model

The TAIDE model, short for Trustworthy AI Dialogue Engine, is an artificial intelligence system developed by the Taiwanese government based on the LLaMA 2 and GPT-3.5 models [1]. It has been primarily optimized for Mandarin Chinese and other Taiwanese dialects. Trained on a vast corpus of textual data, the TAIDE model is capable of generating fluent and coherent textual responses. It demonstrates strong performance in various natural language processing tasks, including article summarization, translation, and dialogue generation.

The model is designed with a strong emphasis on accuracy and ethical compliance, aiming to provide users with reliable and valuable information. Its ability to understand and generate human-like language makes it suitable for a wide range of applications, such as customer service, content creation, and natural language translation.

Currently, TAIDE is a unimodal large language model (LLM) and does not support multimodal data processing. However, future developments will focus on continuous enhancement and expansion to keep pace with evolving technological and societal needs. Multimodal data processing, which involves integrating information from multiple modalities such as text, images, and audio, is a rapidly emerging field with substantial application potential in domains including healthcare, education, and autonomous systems. As research in artificial intelligence progresses, extending the TAIDE model to support multimodal capabilities is considered a critical and timely advancement to enhance its functionality and provide greater value to end users.

1.2.2. Multimodal Learning and Medical Image Diagnosis

With the advancement of artificial intelligence, multimodal learning (ML) has emerged as a prominent research direction in medical image analysis. Traditional image classification models, such as convolutional neural networks (CNNs), have been widely applied in the classification and diagnosis of skin lesions. For example, Swathi et al. (2023) proposed the use of CNN architectures including VGG16, ResUNet, and InceptionV3 for classifying skin lesion images, demonstrating effective classification performance [2].

However, approaches based solely on image classification present certain limitations, particularly in their inability to generate detailed diagnostic descriptions or provide interpretability. As a result, multimodal techniques that integrate both images and textual data are gaining increasing attention. According to a comprehensive study by Zhang et al. (2024), visual language models (VLMs) are capable of learning associations between visual and linguistic information from large-scale multimodal datasets and exhibit strong zero-shot reasoning abilities across various visual recognition tasks [3].

These technologies show great promise in medical imaging applications, particularly in radiographic image analysis and automated report generation, contributing to improved diagnostic accuracy, increased automation, and enhanced efficiency in healthcare workflows. Therefore, the integration of multimodal learning approaches with VLMs is expected to become a key direction in the future development of medical imaging diagnosis.

 doi:10.20944/preprints202604.1897.v1

1.2.3. Visual RAG

With the success of large-scale pre-trained language models (LLMs) in the field of natural language processing, Retrieval-Augmented Generation (RAG) has emerged as a promising technique that combines pre-trained models with non-parametric memory to enhance knowledge retrieval and content generation [4]. RAG dynamically acquires external information through a retrieval mechanism and integrates it into the model's generative process, significantly improving performance in knowledge-intensive tasks such as open-domain question answering.

Recent studies have extended RAG frameworks to multimodal applications by incorporating both visual and textual information to improve the comprehension and generation of mixed-modality documents. For instance, Yu et al. (2024) introduced VisRAG, a model that employs visual language models (VLMs) to embed documents as images and enhances generative capabilities through retrieval, achieving substantial performance improvements [5]. Similarly, Bonomo and Bianco (2025) proposed Visual RAG, which integrates contextual learning with a retrieval mechanism to dynamically select the most relevant examples, thereby enhancing the learning capability of multimodal large language models (MLLMs). Experimental results demonstrate that Visual RAG can achieve performance comparable to multi-example in-context learning (ICL) in image classification tasks while significantly reducing computational overhead [6].

In the context of medical image analysis, RAG presents a novel and promising technological direction by integrating retrieval mechanisms with VLMs to facilitate the automatic generation of diagnostic reports. Although current research primarily focuses on open-domain question answering and text generation tasks, the application of these models in medical imaging is expected to play a critical role in improving diagnostic accuracy, interpretability, and clinical decision-making in the future.

1.2.4. Development of Visual-Language Models (VLMs) in Medical Image Analysis

Vision-Language Models (VLMs), which integrate image processing with natural language generation, have made significant advancements in recent years and demonstrated strong potential in medical image analysis and diagnostic support. These models are capable of understanding visual content and generating semantically rich textual descriptions, thereby improving diagnostic accuracy and the interpretability of medical reports.

For instance, MiniGPT-4 combines a pre-trained Vision Transformer (ViT) with a large language model (LLM), enabling the generation of detailed and contextually relevant descriptions from image content [7]. CheXbert has further demonstrated the applicability of VLMs in the medical domain by generating radiology reports from chest X-ray images [8]. The development of multimodal technologies has been further advanced by MiniGPT-v2, introduced by Chen et al. (2023), a highly adaptable model capable of performing various tasks such as image captioning, visual question answering, and localization [9].

Building on these prior works, the present study extends the application of MiniGPT-v2 by integrating it with TAIDE-LX-7B, a large-scale traditional Chinese language model. This integration aims to fine-tune the generation of diagnostic descriptions for skin lesions, with the goal of enhancing both the readability and accuracy of Chinese-language medical diagnostic reports.

1.2.5. Skin Lesion Diagnostic Aid Systems

In recent years, considerable research has been dedicated to the development of deep learning-based diagnostic systems for skin lesions, among which SkinGPT-4 represents one of the most advanced applications of Visual Large Language Models (VLLMs). SkinGPT-4 employs a fine-tuned version of MiniGPT-4, trained through a two-stage process involving a large collection of skin lesion images, clinical concepts, and physicians' notes. This training strategy enables the model to accurately recognize dermatological features and generate diagnostic outputs [10].

 doi:10.20944/preprints202604.1897.v1

In this study, a multimodal diagnostic system was further developed and deployed on Web Chat and LINE BOT platforms to enhance usability and accessibility in real-world clinical settings. The integration of the system into these communication tools aims to provide users with a more convenient and efficient diagnostic experience, thereby increasing the practicality and adoption of AI-assisted diagnostic technologies in everyday healthcare environments.

2. Materials and Methods

2.1. Data Source

This study used the SkinCAP dataset, which includes 4,000 images of skin lesions along with their corresponding textual descriptions, to train and evaluate the diagnostic performance of a multimodal language model. The images were sourced from the Fitzpatrick 17k Dermatology Dataset and the Diverse Dermatology Imaging Dataset, with annotations provided by board-certified dermatologists to ensure a diverse and clinically relevant range of medical descriptions and visual examples [11].

During the data preprocessing stage, all images were uniformly resized, and the accompanying textual data were translated into traditional Chinese to ensure consistency throughout the model training process.

2.2. Methodology

This study employed two distinct approaches to extend the unimodal TAIDE model for multimodal applications, aiming to integrate both image and textual data to enhance the generation of diagnostic descriptions for skin lesions. The two approaches—Visual Retrieval-Augmented Generation (Visual RAG) and Vision-Language Large Model (VLLM) Fine-Tuning—were comparatively evaluated in terms of their effectiveness.

As illustrated in Figure 1, the left panel presents the workflow of the Visual RAG framework, which involves the following steps: inputting a skin lesion image, extracting features using the EfficientNetV2B2 model, retrieving visually similar reference images, generating semantic prompts based on the retrieved images, and finally feeding these prompts into the TAIDE model to produce diagnostic descriptions. The right panel of Figure 1 illustrates the workflow of the VLLM Fine-Tuning framework. In this approach, the input images are processed through a multimodal fine-tuning pipeline that integrates the MiniGPT-V2 architecture with the TAIDE model. The model learns the alignment between visual features and diagnostic text, enabling it to directly generate diagnostic descriptions in traditional Chinese.

 doi:10.20944/preprints202604.1897.v1

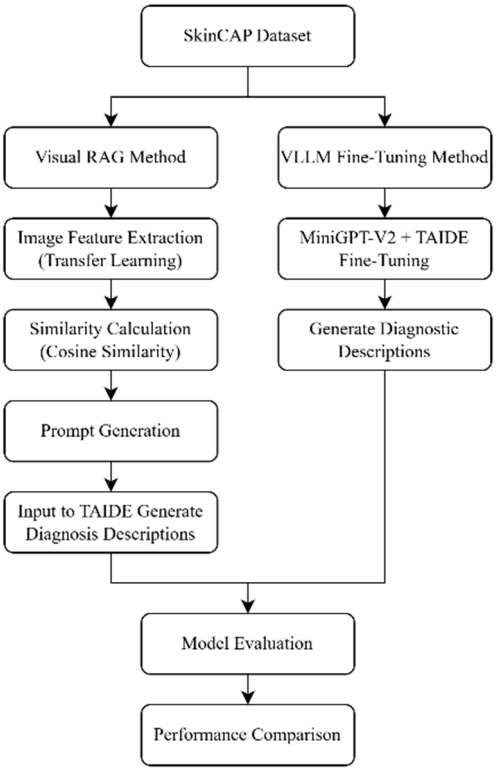

Figure 1. Study Flowchart.

2.2.1. Visual RAG-Based Retrieval with TAIDE for Diagnostic Description Generation

This approach adopts the Visual Retrieval-Augmented Generation (Visual RAG) framework, which integrates content-based image retrieval with language generation to support the automated generation of diagnostic reports. Specifically, the system retrieves semantically and visually similar dermatological images from a reference database to provide contextual grounding for the language model. Two strategies were utilized for image feature extraction and retrieval:

(1) **Pre-trained Feature Extraction**: The EfficientNetV2B2 convolutional neural network, pre-trained on a large-scale image dataset, was employed as a feature extractor. It was applied to 3,200 training images of skin lesions to obtain 1,408-dimensional feature vectors, forming a high-dimensional embedding space for retrieval.

(2) **Fine-Tuned Feature Extraction**: The same EfficientNetV2B2 model was fine-tuned on the 3,200 skin lesion images to adapt its representations to the dermatological domain. The resulting domain-specific embeddings were then used to construct an alternative image feature set.

The feature vectors obtained from both methods were used independently to compute cosine similarity scores against the feature set, enabling the retrieval of the most relevant dermatologic images. Based on the retrieved results, semantic prompts were generated and incorporated into prompts for the TAIDE model to generate diagnostic descriptions.

2.2.2. VLLM Fine-Tuning with MiniGPT-V2 and TAIDE for Multimodal Generation

The second approach adopts a Vision-Language Large Model Fine-Tuning (VLLM Fine-Tuning) strategy, leveraging MiniGPT-V2 as the foundational multimodal architecture. To enhance its capability in generating diagnostic descriptions in traditional Chinese, the model is integrated with TAIDE-LX-7B, a unimodal language model optimized for traditional Chinese text generation.

The combined model is fine-tuned on the SkinCAP dataset to improve its alignment between visual inputs and linguistic outputs, thereby enabling more accurate and contextually appropriate generation of diagnostic descriptions for skin lesions in traditional Chinese.

 doi:10.20944/preprints202604.1897.v1

2.2.3. Evaluation Metrics

To evaluate the quality and accuracy of the generated diagnostic descriptions, five widely adopted automatic evaluation metrics are employed: Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Consensus-based Image Description Evaluation (CIDEr), and Semantic Propositional Image Caption Evaluation (SPICE). These metrics provide a comprehensive assessment of the generated texts in terms of lexical overlap, semantic structure, and agreement with reference descriptions.

Given that the generated diagnostic texts may differ lexically from the reference descriptions while still being clinically valid, low scores on these metrics do not necessarily indicate poor diagnostic quality. To address this limitation and better capture real-world applicability, a supplementary human evaluation was conducted. In this assessment, we rated the diagnostic descriptions on a five-point Likert scale (1 = highly unsatisfactory, 5 = highly satisfactory), providing a more nuanced evaluation of the clinical relevance and linguistic adequacy of the generated content.

2.3. System Development

To support practical deployment, the proposed models are integrated into a web-based chatbot and a LINE BOT service platform, enabling users to upload images of skin lesions and receive corresponding diagnostic descriptions. The system seamlessly combines image analysis and natural language generation capabilities within a user-friendly interface, thereby enhancing accessibility for both healthcare professionals and the general public.

3. Results

3.1. Visual RAG-Based Migration Learning Retrieval Combined with TAIDE Model for Search Generation

The 4,000 images were divided into a training set of 3,200 images and a test set of 800 images, with an 80:20 ratio. Both the pre-trained EfficientNetV2B2 model and the fine-tuned EfficientNetV2B2 model were evaluated as feature extractors, and their performance results are summarized in Table 1.

Based on the evaluation metrics BLEU-1, METEOR, ROUGE-L, CIDEr, and SPICE, the results clearly demonstrate that the fine-tuned EfficientNetV2B2 model outperforms the pre-trained version. This performance improvement can be attributed to the fine-tuning process, which allows the model to adapt to the specific requirements of the task, refine its feature extraction capabilities, and achieve better alignment with the evaluation criteria, thus yielding superior results compared to the pre-trained model.

Table 1. Performance Evaluation of Visual RAG-based Migration Learning Retrieval Combined with TAIDE Model for Search Generation. *The highest score for each evaluation metric is highlighted in bold.*

Feature Extractor	Bleu_1	METEOR	ROUGE_L	CIDEr	SPICE
Pretrained EfficientNetV2B2	0.011	0.039	0.166	0.000	0.133
Fine-tuned EfficientNetV2B2	**0.090**	**0.094**	**0.196**	**0.002**	**0.137**

3.2. VLLM Fine-Tuning Based MiniGPT-V2 Framework Combined with TAIDE Model for Fine-Tuning Generation

This study employed the MiniGPT-V2 framework in conjunction with the TAIDE model to conduct a comprehensive evaluation of model generation performance across various data cut-off

 doi:10.20944/preprints202604.1897.v1

ratios and class selections. Due to the specific requirements of the study, which necessitated the inclusion of both images and professional diagnostic descriptions, self-performed data augmentation was not feasible. As a result, categories with balanced data distributions were selected for testing to assess the model's generation capabilities. The key findings from the analysis are summarized as follows:

(1) **Full Dataset (4000 Images, 179 Categories, 8:2 Split)**: The dataset, consisting of 4,000 images, was partitioned into a training set (80%) and a test set (20%) using an 80:20 split. However, due to the uneven distribution of categories and the limited number of samples in certain disease categories, the model encountered difficulties in effectively learning the features of all categories. This resulted in suboptimal performance and evaluation outcomes.

(2) **Subset Dataset (1061 Images, 15 Categories, 8:2 Split):** Fifteen categories, each containing between 50 and 100 samples (totaling 1,061 images), were selected for testing using the same 80:20 split. While the issue of data imbalance was alleviated, the relatively small sample size within each category limited the model's ability to learn effectively, hindering improvements in classification accuracy.

(3) **Balanced Dataset (515 Images, 4 Categories, 8:2 Split):** Four categories, each with more than 100 samples (totaling 515 images), were chosen for testing with an 80:20 split. At this stage, the problem of data imbalance was resolved, and the sufficient sample size within each category significantly improved the model's diagnostic capabilities for specific disease categories. The test results demonstrated that a balanced data distribution notably enhanced model performance, particularly in the generation of diagnostic descriptions. The CIDEr score reached its highest value, and the model received the highest subjectivity rating in terms of description quality.

Table 2. Performance Evaluation of VLLM Fine-tuning based MiniGPT-V2 framework combined with TAIDE model for fine-tuning generation. *The highest score for each evaluation metric is highlighted in bold.*

Dataset Split (Image Count)	Bleu_1	METEOR	ROUGE_L	CIDEr	SPICE	Subjective Rating
4000 images (8:2 split)	0.3167	0.2171	0.3331	0.205	0.2362	2.5013/5
1061 images (8:2 split)	0.318	0.2166	**0.3376**	0.3413	0.2506	3.3568/5
515 images (8:2 split)	**0.3605**	**0.2232**	0.3369	**0.5672**	**0.3062**	**3.6923/5**

All metric scores reported in Tables 1 and 2 are unitless and were computed using the standard COCO caption evaluation toolkit. Higher values indicate better alignment between the generated and reference descriptions. While most metrics (e.g., BLEU, ROUGE-L, METEOR, and SPICE) are bounded between 0 and 1, CIDEr is not strictly limited to this range and can exceed 1.0 when multiple reference captions are available. In this study, however, only a single reference caption was provided per image, which naturally results in lower CIDEr scores despite the generated outputs potentially being clinically valid.

3.3. System Demonstration

In this study, we implemented Skin-TAIDE: A Multimodal Traditional Chinese Skin Lesion Diagnosis Description Generation System, which allows users to upload skin lesion images via either a Web Chat or LINE BOT interface. Upon image upload, the system automatically processes the image and generates a diagnostic description based on the selected model. Figure 2 presents an example of the Web Chat interface, in which the MiniGPT-V2 framework—fine-tuned using the

 doi:10.20944/preprints202604.1897.v1

VLLM method—is integrated with the TAIDE model to produce diagnostic descriptions in traditional Chinese.

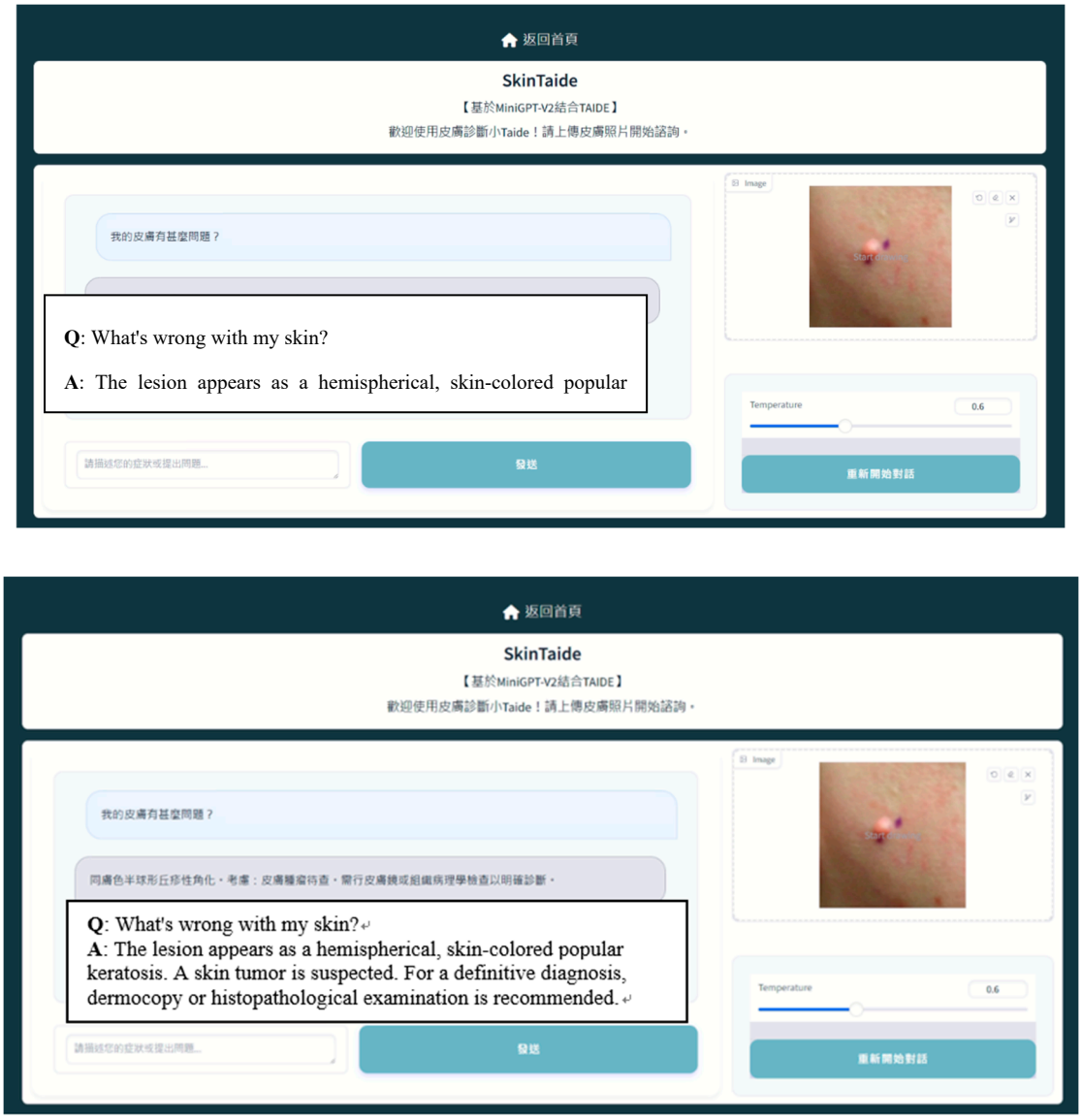

Figure 2. Q&A page of the "Skin-TAIDE" system developed in this study.

4. Discussion

The results indicate that the Visual RAG method extracts image features using a pre-trained CNN model (EfficientNetV2B2) and retrieves the most relevant dermatological images based on feature similarity to generate diagnostic cues. These cues are then integrated with the TAIDE model to generate diagnostic descriptions. While this method successfully captures certain visual features, it is limited by its reliance on the retrieval mechanism, which restricts the diversity and accuracy of the generated descriptions.

In contrast, the VLLM fine-tuning approach employs the MiniGPT-V2 framework as a multimodal backbone to enhance diagnostic description generation through joint learning of image and text modalities. This approach focuses on optimizing the fusion of visual and textual information, thereby improving the accuracy and contextual relevance of the generated outputs. However, both methods are affected by data imbalance and limited dataset size, highlighting the need for further optimization to improve diagnostic precision. The following discussion highlights the distinctions between this study and existing research.

 doi:10.20944/preprints202604.1897.v1

Unlike the original TAIDE model (1), which is a unimodal language model designed for text generation tasks, the Visual RAG and VLLM fine-tuning approaches extend TAIDE into a multimodal framework by incorporating image data. As the TAIDE model lacks the inherent ability to process and interpret visual inputs, our proposed framework enhances diagnostic generation by integrating visual cues, thereby improving contextual understanding and diagnostic accuracy.

The Visual RAG method (6) differs from our approach in both image retrieval and diagnostic description generation. Visual RAG is primarily designed for Visual Knowledge-Intensive Question Answering (VKIQ), leveraging CLIP for text-to-image retrieval and exploring how Multimodal Large Language Models (MLLMs) can extract information directly from images to augment responses. In contrast, our approach is tailored to medical imaging diagnosis, using EfficientNetV2B2 for image feature extraction and similarity matching. Retrieved images are used to generate prompts, which are then combined with the TAIDE model to produce diagnostic descriptions.

MiniGPT-V2 (9) serves as a robust multimodal foundation model. In this study, it is fine-tuned specifically for generating diagnostic descriptions of skin conditions in traditional Chinese. While the original MiniGPT-V2 was trained for general-purpose multimodal tasks, our fine-tuned version incorporates domain-specific dermatological data, enabling it to generate more accurate and contextually relevant diagnostic descriptions. This demonstrates the critical importance of domain-specific adaptation when applying general models to specialized medical tasks.

SkinGPT-4 (10), considered one of the most advanced models for dermatological diagnosis, is a refined version of MiniGPT-4. It benefits from extensive multimodal pretraining and a substantially larger dataset, enhancing its performance in diagnosing skin lesions in English. By comparison, our VLLM fine-tuned approach is built on MiniGPT-V2 and specifically trained on traditional Chinese dermatological texts, making it more effective for Chinese-language diagnostic tasks. Furthermore, this study integrates both the Visual RAG and VLLM fine-tuning methods, providing complementary perspectives and enriching the diversity of approaches for improving diagnostic description generation.

5. Conclusions

This study investigates two approaches for extending a unimodal model to multimodal applications for generating diagnostic descriptions of skin lesions: (1) Visual RAG-based transfer learning retrieval integrated with the TAIDE model, and (2) VLLM fine-tuning based on the MiniGPT-V2 framework, also integrated with the TAIDE model. In addition, we developed a skin lesion diagnostic description generation system to demonstrate the potential of multimodal models in supporting clinical decision-making.

Overall, this study validates the effectiveness of both the Visual RAG and VLLM fine-tuning methods in the context of skin lesion diagnosis, highlighting the feasibility of extending unimodal models to multimodal tasks. Future improvements in diagnostic performance and clinical applicability are anticipated if the challenges of data distribution imbalance can be addressed and a larger, more diverse dataset is incorporated.

Funding: This study was partial funded by the National Science and Technology Council, Taiwan, R.O.C., grant number: NSTC 113-2121-M-040-002.

Acknowledgments: This study would like to thank three students, Chen Zhi-Ling, Huang Yi-Ting, and Huang Qin-Yong, for their assistance in data processing and system testing.

References

1. (NARLabs) STPRaIC. Trustworthy AI Dialogue Engine (TAIDE) 2024 [Available from: https://taide.tw].
2. Swathi B, Kannan K, Chakravarthi SS, Ruthvik G, Avanija J, Reddy CCM, editors. Skin cancer detection using VGG16, InceptionV3 and ResUNet. 2023 4th International Conference on electronics and sustainable communication systems (ICESC); 2023: IEEE.

 doi:10.20944/preprints202604.1897.v1

3. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024.
4. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems. 2020;33:9459-74.
5. Yu S, Tang C, Xu B, Cui J, Ran J, Yan Y, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. arXiv preprint arXiv:241010594. 2024.
6. Bonomo M, Bianco S. Visual RAG: Expanding MLLM visual knowledge without fine-tuning. arXiv preprint arXiv:250110834. 2025.
7. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:230410592. 2023.
8. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:200409167. 2020.
9. Chen J, Zhu D, Shen X, Li X, Liu Z, Zhang P, et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:231009478. 2023.
10. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. SkinGPT-4: an interactive dermatology diagnostic system with visual large language model. arXiv preprint arXiv:230410691. 2023.
11. Zhou J, Sun L, Xu Y, Liu W, Afvari S, Han Z, et al. SkinCAP: A Multi-modal Dermatology Dataset Annotated with Rich Medical Captions. arXiv preprint arXiv:240518004. 2024.