WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 4, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n4.01
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Self-Evolving Diagnostic Framework based Gated Residual Adapters and OpenClaw-Based Medical Agents

Shuai Feng, Pan Su

School of Control and Computer Engineering, North China Electric Power University, Baoding, 071003, China

ABSTRACT

This Accurate and interpretable report generation from fundus images remains a critical yet challenging task in medical artificial intelligence, particularly due to the static nature of model adaptation and the limited evolvability of existing agent-based frameworks. Although ophthalmic foundation models have significantly improved visual representation learning through large-scale self-supervised pretraining, they lack mechanisms for continual adaptation during inference. Meanwhile, current agent-based approaches enhance reasoning but remain constrained by fixed cognitive structures. In this work, we propose a self-evolving diagnostic framework that unifies parametric adaptation and cognitive evolution for fundus report generation. Specifically, we introduce a gated residual adapter to enable dynamic, inference-time knowledge integration while preserving prior knowledge. Furthermore, we develop a medical agent architecture based on the OpenClaw paradigm, which continuously refines its core reasoning strategy through physician feedback and consistency-driven constraints. By coupling model-level adaptability with agent-level reasoning evolution, the proposed framework enables sustained performance improvement in complex clinical scenarios. This work provides a new perspective on building continuously evolving intelligent diagnostic systems for real-world healthcare applications.

KEYWORDS

Fundus Image Analysis; Gated Residual Adapter; Self-Evolving Framework; Medical Agent

1. INTRODUCTION

The eye is one of the most important sensory organs of the human body and is also highly susceptible to disease. Many retinal disorders remain clinically silent in their early stages, although pathological changes may already be gradually manifested in fundus images. Therefore, efficient and reliable fundus imaging examinations are of critical importance for the early screening and intervention of such diseases. In current clinical practice, color fundus photography (CFP), as one of the most commonly used fundus imaging modalities, has been widely employed in routine physical examinations and large-scale disease screening because of its noninvasive nature, low cost, and operational convenience. CFP enables the intuitive visualization of the overall anatomical structure of the retina, pigment distribution, and the morphological characteristics of major lesions, thereby providing an important data foundation for intelligent diagnosis [1].

With the advent of the large-model era in deep learning, self-supervised pretraining based on large-scale unlabeled data has gradually evolved into a core paradigm for representation learning in medical imaging. Ophthalmic foundation models represented by RETFound [2] have substantially improved the ability of models to capture retinal structures and pathological features through self-supervised reconstruction learning on millions of fundus images.

Relevant studies have shown that such foundation models exhibit better generalization performance than conventional supervised learning models across a variety of downstream tasks and can achieve efficient transfer under few-shot settings [2, 3]. Against this backdrop, the “pretraining-finetuning” paradigm has become the mainstream approach. Large-scale foundation models are increasingly assuming the role of universal representation learners, while retraining backbone models for specific tasks is not only computationally expensive but also unnecessary in most scenarios.

On the other hand, medical report generation, as an important research direction in medical artificial intelligence, aims to transform multimodal clinical data into structured, semantically rigorous, and clinically compliant natural language descriptions. Unlike general image captioning tasks, medical report generation requires not only highly accurate visual understanding but also adherence to strict clinical reasoning processes, as well as the generation of long-form text with multi-level structure. These characteristics render traditional end-to-end generation methods based on a single model significantly challenged in complex clinical scenarios [4]. Accordingly, how to perform efficient adaptation and continual optimization on top of existing foundation models has become a key issue in current research.

In recent years, with the development of large language models, agent-based report generation paradigms have gradually emerged. By introducing multi-module collaborative mechanisms, such approaches decompose the complex process of report generation into multiple subtasks, including perception, reasoning, verification, and expression, and realize decision-making workflows that better conform to clinical logic through chain-of-thought reasoning and tool invocation [5]. Agent-driven generation frameworks not only substantially improve the interpretability and consistency of report content, but also mitigate, to some extent, the hallucination issues commonly observed in conventional generation models, and have therefore become an important direction in medical report generation [6].

Although the development of foundation models and agent technologies has opened new possibilities for medical diagnostic report generation, existing methods still exhibit notable limitations. On the one hand, foundation models typically adopt static fine-tuning strategies in downstream tasks and therefore lack the ability to adapt dynamically during inference. On the other hand, although the reasoning strategies of agents can be optimized through prompt engineering, their core cognitive structures remain relatively fixed, making it difficult to achieve true continual evolution.

To address the above issues, this study proposes a self-evolving diagnostic framework that integrates gated residual adapters with the medical OpenClaw agent paradigm.

At the model level, a gated residual adapter is introduced. By introducing a learnable gating mechanism, the model can dynamically adjust the adapter during inference, thereby continuously incorporating new knowledge while stably preserving prior knowledge, and ultimately improving its capacity for long-term evolution.

At the agent level, a medical scenario-oriented agent architecture based on the OpenClaw paradigm is constructed. Through the continual updating of the soul of agent, the agent can progressively optimize its diagnostic logic under the guidance of physician feedback and consistency constraints, thereby enabling adaptive evolution in the report generation process.

Through this design, the present study unifies the parametric evolution of the model and the cognitive evolution of the agent within a single framework, thereby establishing a diagnostic system capable of jointly evolving representational capacity and reasoning ability. This framework provides a new research paradigm for long-term, sustainable intelligent diagnosis in real-world clinical settings.

2. METHODLOGY

2.1. Multilayer Gated Residual Adapter

To address the lack of dynamic fine-tuning capability in diagnostic models during inference, this study proposes an adapter based on a multilayer gated residual architecture, enabling adaptive parameter adjustment at the inference stage. Through multilayer stacking and a learnable gating mechanism, the adapter dynamically modulates features at different hierarchical levels, thereby achieving effective intra-modal and inter-modal knowledge transfer and fusion while preserving the stability of the original knowledge [7, 8].

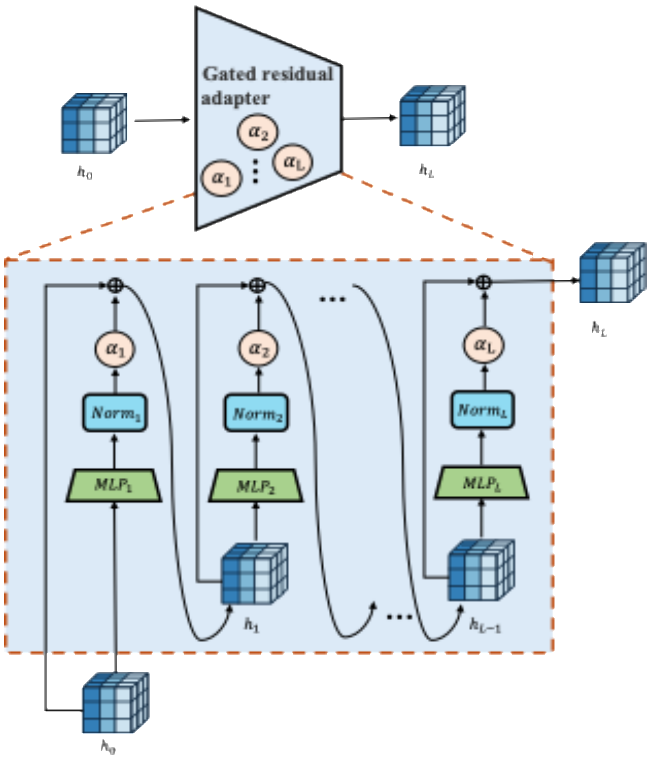

Figure 1. Multilayer Gated Residual Adapter

The adapter is constructed based on a multilayer gated residual architecture, in which learnable gating parameters are introduced to modulate the contribution of the adapter [9]. When no further adaptation is required for a given task, the gating values tend toward zero, effectively suppressing the adapter branch. Due to the residual formulation, the module then collapses into an identity mapping, acting as a no-op that directly preserves the backbone features [10, 11]. This mechanism prevents negative transfer by ensuring that uninformative or detrimental task-specific updates do not degrade the performance of the pretrained model. The overall architecture of the adapter is shown in Fig. 1.

Each adapter consists of a bottleneck-structured multilayer perceptron (MLP). By stacking multiple MLPs, a deep adapter is constructed, while learnable gating parameters are introduced to precisely regulate the contribution of each adapter layer. The specific computational process is formulated as follows:

$$u^{(l)} = LN_l\Big(MLP_l(h^{l-1})\Big), l = 1, \dots, L \tag{1}$$

$$h^{(l)} = h^{(l-1)} + \alpha_l \times u^{(l)}, l = 1, \dots, L \tag{2}$$

Where, $h^{(0)}$ denotes the input feature, LN_l represents the normalization operation at the l-th layer, MLP_l denotes the perceptron at the l-th layer, $u^{(l)}$ is the output of the l-th layer, and α_l is the

learnable gating parameter of the l-th layer. When α_l approaches zero, that layer degenerates into an identity mapping, thereby ensuring that the model does not introduce undesired shifts when no adaptation information is required.

By treating the multilayer structure as a whole, the adapter with a multi-layer gated residual architecture can be concisely expressed as:

$$h^{(L)} = Adapter(h^{(0)}) \tag{3}$$

2.2. Fine-Grained Knowledge Training

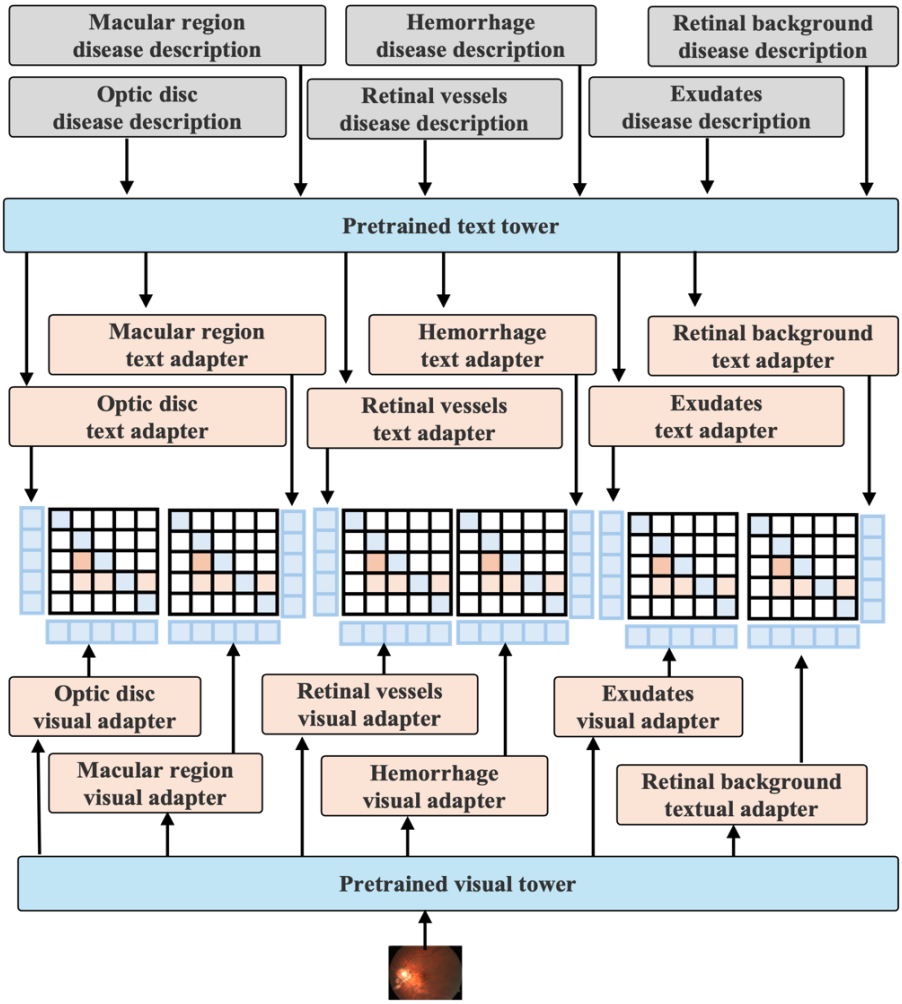

Figure 2. Fine-Grained Knowledge Training

CFP is a static retinal imaging modality based on visible-light imaging. Its principal advantage lies in its ability to intuitively depict the anatomical structure of the retina and the morphological characteristics of lesions, with a stronger emphasis on spatial structure [12-14]. Based on differences in tissue morphology and lesion distribution patterns, the CFP modality focuses on six key morphological features:

Optic disc: The convergence point of retinal nerve fibers with stable anatomical features in CFP. Variations in its morphology, boundary, and color are closely associated with optic nerve-related diseases and provide a key reference for global structural analysis.

Macular region: The central area responsible for visual function, exhibiting characteristic structures in CFP. It is highly susceptible to vision-threatening diseases, and its alterations often serve as early diagnostic indicators.

Retinal vessels: CFP reveals vessel trajectory, branching, and caliber.

Hemorrhage: Appearing as dark patchy or punctate lesions, hemorrhages are closely related to disease type and severity, serving as highly discriminative features.

Exudates: Bright yellow or white deposits associated with vascular leakage, with distinct color and texture that facilitate reliable identification.

Retinal background: Reflects overall retinal condition; variations in color and texture may indicate chronic pathological changes such as ischemia or pigmentation abnormalities.

To achieve aspect-level fine-grained vision-text alignment Training for disease representation, this study introduces aspect-level adapters on both the visual and textual sides.As shown in Fig. 2,the parameters of the pretrained visual tower, namely the large-scale vision model for ophthalmic medicine, and the pretrained text tower, namely the large-scale language model for ophthalmic medicine, are kept frozen, while multiple aspect-level adapters are inserted in parallel. Each adapter adopts a gated residual architecture, enabling fine-grained modulation of semantic features corresponding to different disease aspects while preserving the stability of the backbone networks.

Specifically, on the textual side, multi-aspect semantic descriptions $l_n^{(m)}$ are encoded by the frozen textual tower f_{text} to generate base embeddings:

$$t_n^{(m)} = f_{text}\left(l_n^{(m)}\right) \tag{4}$$

Subsequently, an independent textual adapter $Adapter_{text}^{(m)}$ is assigned to each aspect to obtain phase-level fine-grained textual features:

$$\widehat{t_n^{(m)}} = Adapter_{text}^{(m)}\left(t_n^{(m)}\right) \tag{5}$$

Where, N denotes the total number of diseases, $t_n^{(m)}$ represents the encoded feature of the n-th disease under the m-th aspect produced by the text tower, and $\hat{t}_n^{(m)}$ denotes the output feature obtained after this representation has been enhanced by the text adapter corresponding to the m-th aspect.

On the visual side, the input image of the CFP modality x is fed into the frozen image tower encoder to extract global features:

$$v = f_{img}(x) \tag{6}$$

Subsequently, for the six key aspects in the CFP modality, namely the optic disc, macular region, blood vessels, hemorrhage, exudates, and retinal background, six corresponding visual adapter blocks, $\text{Adapter}_{\text{image}}^{(m)}$, are introduced. Each adapter independently generates the feature representation for the corresponding semantic aspect m:

$$\widehat{v^{(m)}} = Adapter_{image}^{(m)}\ (v) \tag{7}$$

Subsequently, bidirectional contrastive learning is performed using the multi-aspect contrastive matrix $G_B^{(m)}$:

$$\mathcal{L}_{con}^{(m)} = CE\left(\boldsymbol{G_B^{(m)}}, \boldsymbol{S_{v2t}^{(m)}}\right) + CE\left(\boldsymbol{G_B^{(m)^T}}, \boldsymbol{S_{t2v}^{(m)}}\right) \tag{8}$$

Where CE represents the InfoNCE loss, m indexes different fine-grained phase-level aspects, $S_{v2t}^{(m)} = \tau\left(\hat{v}^{(c)}, \hat{t}_n^{(c)}\right)$ and $S_{t2v}^{(m)} = \tau\left(\hat{t}_n^{(c)}, \hat{v}^{(c)}\right)$ represent the phase-level similarity matrices from image to text and from text to image, respectively, τ is a learnable temperature parameter.

2.3. Multi-Agent Collaborative Diagnostic Architecture

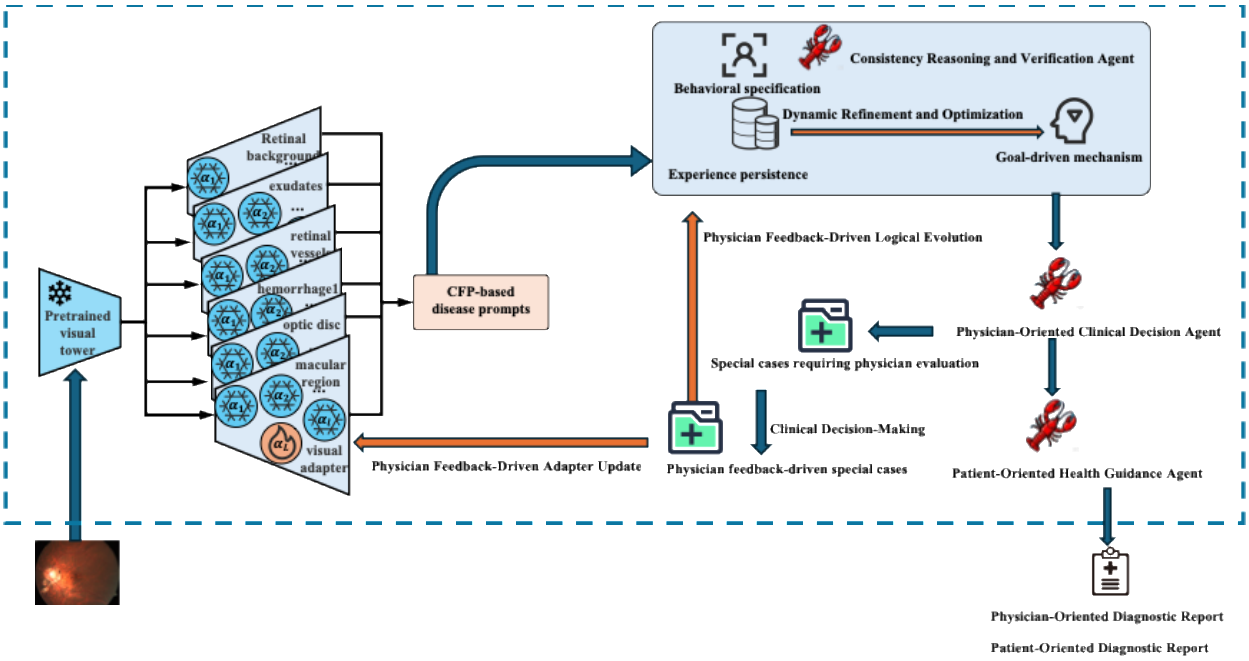

Figure 3. Multi-Agent Collaborative Diagnostic Architecture

Fine-Grained Knowledge Training enables the learning of stable visual representations with explicit spatial structure in the CFP modality, thereby substantially improving the model's performance in ophthalmic disease recognition and providing a solid foundation for high-level semantic understanding of complex fundus images.

However, in real-world clinical settings, relying solely on a single deep model to directly output diagnostic conclusions still presents certain limitations. As shown in Fig. 3, our paper introduces Multi-Agent Collaborative Diagnostic on top of Fine-Grained Knowledge Training and constructs an ophthalmic diagnostic report generation system based on multi-agent collaboration and continual knowledge evolution. The overall framework of the system is and mainly consists of four core modules: the model inference layer, the Consistency Reasoning and Verification Agent, the Physician-Oriented Clinical Decision Agent, and the Patient-Oriented Health Guidance Agent.

As shown on the left side of Fig. 3, during the clinical perception stage, the pretrained model is first employed to perform joint analytical inference on CFP, extracting multiple disease-related semantic features from structural regions across different modalities. Subsequently, a predefined medical knowledge dictionary is introduced to semantically expand the disease categories identified by the model, and CFP-based disease prompts are constructed accordingly, thereby providing unified input information with medical semantic constraints for subsequent agent reasoning and diagnostic report generation.

The Consistency Reasoning and Verification Agent performs consistency analysis on the disease prompts. When conflicts arise among diagnostic results across different disease aspects, the system triggers an anomaly detection mechanism and labels the corresponding cases as special cases. After physician feedback is obtained, these special cases are used as key samples to drive the continual updating and optimization of the model.

Subsequently, in response to the needs of different user groups, differentiated agents are constructed to generate multi-level diagnostic reports. One type is the Physician-Oriented Clinical Decision Agent, whose reports focus on presenting disease-related evidence, key pathological manifestations, and potential risk factors, thereby providing reference support for clinical diagnosis and treatment decision-making. The other type is the Patient-Oriented Health Guidance Agent, whose generated content places greater emphasis on explaining disease conditions in a clear and understandable manner and on providing corresponding health education, daily prevention, and care recommendations, thereby helping patients better understand diagnostic results and improve their self-management capacity.

The Physician-Oriented Clinical Decision Agent submits special cases to physicians for adjudication, and the system then performs targeted dynamic updates to different modules based on physician feedback. Through this feedback-driven dynamic updating mechanism, the system is able to continuously optimize both the reasoning strategies of the agents and the diagnostic capabilities of the professional model, thereby progressively improving overall reasoning performance and report generation quality, and ultimately realizing a closed-loop optimization process from model prediction to the continual evolution of clinical knowledge.

2.3.1. Physician Evaluation of Special Cases

The Consistency Reasoning and Verification Agent first evaluates the reliability of CFP-based disease prompts by analyzing the conflict information produced by the consistency verification module. When conflicts are detected, this indicates the presence of semantic inconsistency or diagnostic uncertainty across different modalities. In such cases, the system integrates disease prompts with conflict signals and performs consistency reasoning and conflict identification to explicitly highlight potential semantic discrepancies and violations of medical logic, thereby providing diagnostic risk alerts for physicians.

Upon receiving these risk alerts, physicians further examine the imaging evidence to make informed decisions. If a definitive diagnosis cannot be established based on the available imaging data, it suggests that the current evidence is insufficient to support a reliable conclusion. In this scenario, the system generates targeted clinical recommendations and suggests additional examinations based on the identified conflicts, producing a corresponding advisory report to assist clinicians in acquiring more discriminative diagnostic information, thereby improving the accuracy of subsequent diagnosis.

Conversely, if physicians are able to determine a definitive diagnosis from the imaging data, the system updates the corresponding disease information and records the case as a special case. The updated knowledge is then fed back into the system to support subsequent reasoning and decision-making processes, enabling dynamic refinement and continual optimization of the model.

When no conflicts are detected during consistency verification, or when consistency is achieved after updating disease information, the CFP-based disease prompts are considered to be logically consistent from a medical perspective. Subsequently, multiple agents are employed to generate multi-level diagnostic reports, providing both professional clinical insights and patient-oriented health guidance.

2.3.2. Physician Feedback-Driven Logical Evolution

With regard to agent system design, OpenClaw has recently emerged as a novel open-source agent paradigm. Its core innovation lies in the design principle of treating files as cognitive carriers. Through structured Markdown files, the framework explicitly models the agent's identity, behavior, and memory, thereby endowing its cognitive structure with interpretability and editability. As shown in Fig. 3, OpenClaw decomposes agent capabilities into three key components:

Behavioral specification: Through the structured configuration file IDENTIFY.md, the agent’s clinical diagnostic criteria, reasoning biases, and interaction paradigms are explicitly defined, thereby providing the model with preset constraints associated with a professional identity.

Experience persistence: By leveraging MEMORY.md together with a hierarchical logging system, clinical experience is stored in a non-parametric manner, while vector-space indexing techniques enable the retrieval of historical cases and extraction of knowledge based on semantic similarity.

Goal-driven mechanism: Through the dynamic loading of the core objective file SOUL.md and the historical experience trajectories stored in MEMORY.md, a closed-loop self-calibration mechanism is established.

Through this design, the agent is no longer confined to a static objective definition, but is instead able to extract clinical feedback from long-term memory and correct logical biases in the reasoning path in real time. We further extend the agent framework by incorporating a physician-in-the-loop special case evolution mechanism. Specifically, special cases identified during the Physician Evaluation of Special Cases stage are persistently stored in MEMORY as structured clinical experience.

To fully leverage these accumulated cases, we introduce a Dynamic Refinement and Optimization process. At predefined intervals, the system performs periodic summarization and abstraction over the stored special cases. Through this process, redundant or low-informative instances are filtered out, while representative patterns, critical diagnostic cues, and recurring sources of inconsistency are distilled into higher-level knowledge representations. This refined knowledge is subsequently written into SOUL, which functions as the core objective and reasoning guidance file of the agent.

In parallel, IDENTIFY defines the system-level prompts that govern the agent's operational behavior, including diagnostic preferences, reasoning constraints, and interaction protocols. Unlike static prompt engineering, this design enables IDENTIFY to serve as a stable behavioral specification, while SOUL evolves dynamically based on accumulated clinical experience.

As the number of special cases increases, the iterative interaction between MEMORY and SOUL establishes a continual knowledge evolution loop. Newly observed cases enrich MEMORY, periodic refinement updates SOUL, and the updated SOUL in turn reshapes the agent's reasoning trajectory and decision boundaries. In this manner, the agent is progressively adapted to complex and long-tail clinical scenarios, enabling more robust, consistent, and knowledge-aware diagnostic behavior over time.

2.3.3. Physician Feedback-Driven Adapter Update

The text encoder and image encoder, having undergone large-scale pretraining, possess strong general feature representation capabilities. In contrast, the adapter module is designed to capture fine-grained disease-specific knowledge through targeted training, thereby enhancing the discriminability of disease representations.

During the Physician Feedback-Driven Adapter Update stage, the internal MLP parameters within the adapter remain fully frozen. In addition, all gating parameters are kept frozen except for the final gating parameter. This design ensures that the knowledge acquired in previous training stages is maximally preserved, preventing catastrophic forgetting. As shown in Fig. 3, the frozen parameters from the previous stage are denoted by blue snowflake markers, while the unfrozen components are indicated by red flame markers.

Specifically, special cases identified through physician feedback are continuously accumulated and used as high-value supervision signals. Instead of updating the entire adapter, only the final gating parameter is selectively optimized based on these cases. This parameter acts as a high-level controller that modulates the contribution of learned features, allowing the model to dynamically adjust its decision boundaries in response to challenging or ambiguous scenarios.

As more special cases are incorporated, the final gating parameter is iteratively refined, enabling the adapter to gradually evolve without disrupting previously learned representations. This lightweight yet effective update strategy establishes a form of automatic evolution, where the model continuously adapts to new clinical patterns while maintaining stability. In this way, the system achieves a balance

between knowledge retention and adaptive plasticity, leading to improved robustness in handling long-tail and complex diagnostic cases.

3. EXPERIMENTS

3.1. Fine-Grained Knowledge Training

Pre-training Data: All experiments were conducted on a private clinical CFP dataset collected and annotated by experienced ophthalmologists during routine diagnosis. Each sample corresponds to a single examined eye and multi-label disease annotations. The dataset contains 14,870 eyes in total, with 10,387 for training, 2,258 for validation, and 2,225 for testing.

Optimizer and Training Strategy: We use the AdamW [15] optimizer with a cosine annealing learning rate schedule [16]. Linear warm-up [17], periodic decay, and a final cooling stage are applied to stabilize optimization. To better optimize different trainable components, we adopt grouped learning rates, assigning separate learning rates to different parameter groups.

Aspect-Level Contrastive Learning: RetFound [2] is adopted as the visual tower due to its strong CFP representation capability, while BioClinicalBERT [18] is used as the textual tower to model medical semantics in angiographic reports. The parameters of visual tower and textual tower are frozen throughout training. The aspect-level adapters adopt the multi-layer gated residual adapter. Only the aspect-level visual adapters, aspect-level textual adapters, and the temperature parameter in contrastive learning are optimized.

key hyperparameters: The aspect-level adapter adopts a multi-layer gated residual architecture, where its core component consists of a stack of bottleneck MLPs Given an input feature of dimension d, the adapter first projects the feature into a lower-dimensional space of size d/r, followed by a nonlinear transformation, and then restores it back to the original dimension. To flexibly control the capacity and expressiveness of the adapter, two key hyperparameters are introduced. n-the number of bottleneck MLP layers within the residual adapter, which determines the depth of semantic injection; r-the dimensionality reduction ratio, which controls the parameter scale and representational capacity of the adapter. The experimental results are shown in Table 1.

Table 1. Performance Comparison of Different Adapters in fine-grained knowledge training

Method	Number of parameters	AUC/%
r16-n1	75,265	90.453
r16-n3	222,723	92.891
r16-n6	443,907	93.702
r32-n1	38,401	87.847
r32-n3	112,131	92.295
r32-n6	222,723	93.241
r64-n1	19,969	81.765
r64-n3	56,835	85.794
r64-n6	112,131	92.523

These results demonstrate that increasing the adapter depth enhances the model's ability to capture complex disease semantics, while larger adapter capacity further improves alignment performance at the expense of computational efficiency.

A comparative analysis of the reduction ratio r and the number of layers n in the aspect-level adapter reveals that the configuration (r = 16, n = 6) achieves the highest AUC performance (93.702%). However, this setting incurs a substantially larger number of parameters. In contrast, the configuration (r = 32, n = 6) attains a competitive AUC score (93.241%) while maintaining a parameter scale

comparable to that of (r = 16, n = 3), demonstrating a more favorable trade-off between performance and parameter efficiency.

Therefore, considering both effectiveness and efficiency, the configuration (r = 32, n = 6) is adopted as the final setting for the aspect-level visual adapter in the CFP modality during subsequent inference.

3.2. Physician Feedback-Driven Logical Evolution And Adapter Update

To quantitatively evaluate the efficacy of the proposed self-evolving framework, this study employs the number of Special Cases Requiring Physician Evaluation as the primary performance metric. When a special case triggers the need for clinical decision-making, we simulate physician feedback based on the corresponding original diagnostic report. During evaluation, the test set is processed in an iterative manner. After every 100 model iterations, we record the number of special cases detected and apply physician feedback-driven logical evolution and adapter update. The experimental results are shown in Fig. 4.

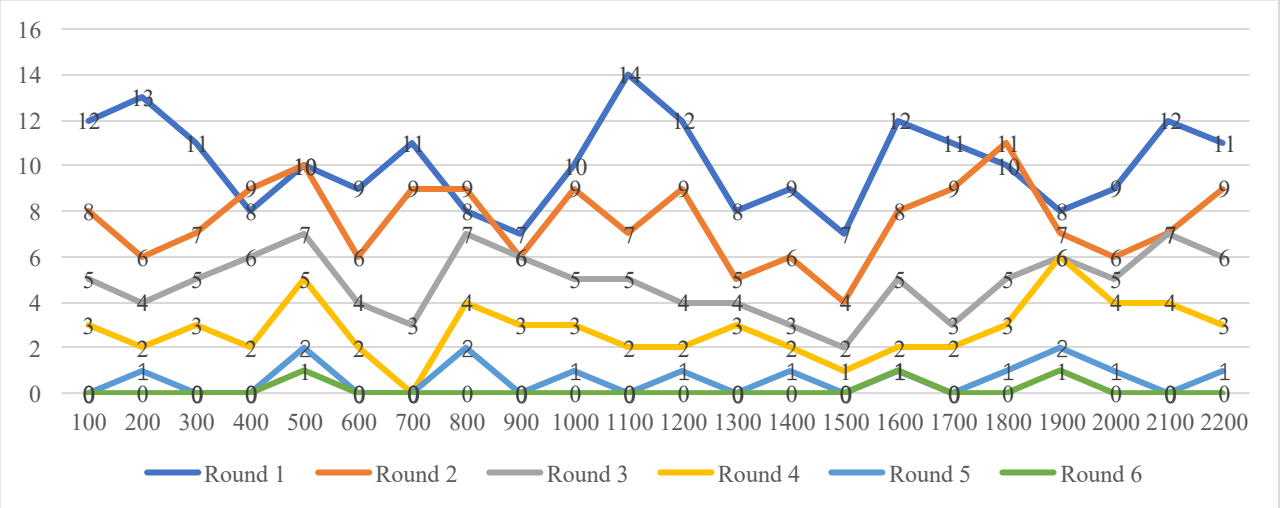

Figure 4. Evolution of Conflict Case Counts Over Iterative Physician Feedback-Driven Updates

As shown in Fig. 4, the results exhibit a clear downward trend in the number of conflict cases as the number of evolutionary rounds increases. In the early rounds (Round 1 and Round 2), conflict case counts remain relatively high across most evaluation points. However, with the progression of evolution — driven by physician feedback and the combined logical and adapter updates — a consistent reduction in conflict case counts is observed in subsequent rounds. By Round 5 and Round 6, most checkpoints show conflict case counts approaching or reaching zero. This progressive decline demonstrates that the proposed physician feedback-driven evolution framework effectively reduces logical inconsistencies identified by the verification agent.

4. CONCLUSION

This paper proposes a self-evolving framework for CFP-based diagnostic report generation that unifies parameter-efficient adaptation via gated residual adapters with logical evolution of OpenClaw-based medical agents. Experimental results demonstrate that the adapter effectively captures fine-grained disease semantics, and the physician feedback-driven updates steadily reduce conflict cases, enhancing reasoning consistency and clinical usability. These findings validate the feasibility of combining lightweight adaptation with continual agent-level optimization for intelligent ophthalmic diagnosis. Future work will extend the evaluation with richer diagnostic metrics, broader physician-in-the-loop studies, and multi-center clinical datasets to further examine generalizability and real-world applicability.

REFERENCES

[1] Santos, A. R., Lopes, M., Santos, T., Reste-Ferreira, D., Marques, I. P., Yamaguchi, T. C., ... & Cunha-Vaz, J. (2024). Intraretinal microvascular abnormalities in eyes with advanced stages of nonproliferative diabetic retinopathy: comparison between UWF-FFA, CFP, and OCTA—the RICHARD study. Ophthalmology and Therapy, 13(12), 3161-3173.

[2] Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., ... & Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. Nature, 622(7981), 156-163.

[3] Shurrab, S., & Duwairi, R. (2022). Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Computer Science, 8, e1045.

[4] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., ... & Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36, 28541-28564.

[5] Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., ... & Gerstein, M. (2024, August). Medagents: Large language models as collaborators for zero-shot medical reasoning. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 599-621).

[6] Chen, K., Qi, J., Huo, J., Tian, P., Meng, F., Yang, X., & Gao, Y. (2025, April). A self-evolving framework for multi-agent medical consultation based on large language models. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[7] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In International conference on machine learning (pp. 2790-2799). PMLR.

[8] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., & Gurevych, I. (2021, April). Adapterfusion: Non-destructive task composition for transfer learning. In Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume (pp. 487-503).

[9] Rebuffi, S. A., Bilen, H., & Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. Advances in neural information processing systems, 30.

[10] Savarese, P., & Figueiredo, D. (2017). Residual gates: A simple mechanism for improved network optimization. In Proc. Int. Conf. Learn. Representations.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[12] Besenczi, R., Tóth, J., & Hajdu, A. (2016). A review on automatic analysis techniques for color fundus photographs. Computational and structural biotechnology journal, 14, 371-384.

[13] Kinouchi, R., Ishiko, S., Hanada, K., Hayashi, H., Mikami, D., & Yoshida, A. (2021). Identification of risk factors for retinal vascular events in a population-based cross-sectional study in Rumoi, Japan. Scientific Reports, 11(1), 6340.

[14] Yang, W. H., Xu, Y. W., & Sun, X. H. (2025). Guidelines for glaucoma imaging classification, annotation, and quality control for artificial intelligence applications. International Journal of Ophthalmology, 18(7), 1181.

[15] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[16] Johnson, O. V., Xinying, C., Khaw, K. W., & Lee, M. H. (2023). ps-CALR: periodic-shift cosine annealing learning rate for deep neural networks. IEEE access, 11, 139171-139186.

[17] Kalra, D. S., & Barkeshli, M. (2024). Why warmup the learning rate? underlying mechanisms and improvements. Advances in Neural Information Processing Systems, 37, 111760-111801.

[18] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.