

RESEARCH



Medrecord-CLIP: enhancing fundus disease diagnosis via EHR-guided vision-language pre-training

Lei Shi^{1†}, Wenbin Zhai^{1†}, Lei Yu^{3*}, Daoying Wen¹, Jing Wang⁴, Shuwei Qiu² and Cuiyun Gao²

Abstract

Automated analysis of Color Fundus Photography (CFP) is essential for large-scale retinal disease screening. However, conventional vision-only models often rely on rigid categorical labels, neglecting the rich clinical nuances found in medical narratives. While existing Vision-Language Pre-training (VLP) frameworks have explored text-based supervision, they frequently overlook individualized patient contexts within Electronic Health Records (EHRs). To bridge this gap, we construct **MedRecordFundus**, a large-scale multimodal dataset pairing 21,290 CFP images with expert-curated EHR narratives across four clinical dimensions. Leveraging this resource, we propose **MedRecord-CLIP**, a knowledge-enhanced foundation model featuring a diagnosis-guided cross-attention mechanism to adaptively extract and fuse salient patient history with diagnostic representations. To improve representation robustness against highly similar clinical descriptions, we introduce a symmetric Multi-Positive InfoNCE objective. Extensive experiments on four public benchmarks—RFMiD, ODIR, APTOS 2019, and IDRiD—demonstrate that MedRecord-CLIP yields statistically significant improvements over state-of-the-art baselines like RETFound and FLAIR on the multi-disease benchmarks (RFMiD and ODIR), and also leads on the IDRiD grading benchmark, while performing on par with the strongest baseline on APTOS 2019. Our approach highlights the critical value of integrating personalized clinical context to enhance the generalizability and interpretability of fundus foundation models.

Keywords: Multimodal learning, Vision-language pre-training, Fundus imaging, Electronic health records

Introduction

Retinal diseases, including diabetic retinopathy and age-related macular degeneration, remain leading causes of preventable visual impairment worldwide [1, 2]. Color fundus photography (CFP) has emerged as the primary modality for population-scale retinal screening, owing to its non-invasive nature and cost-effectiveness [3, 4]. While recent deep learning advances have demonstrated impressive performance in automated CFP analysis, conventional vision-only models are fundamentally constrained by coarse-grained supervision signals, typically one-hot encoded labels, which impose a semantic

bottleneck by discarding the rich linguistic descriptions that link visual cues to clinical reasoning [5–7]. This bottleneck limits their semantic granularity and representational robustness, particularly in scenarios involving early-stage disease detection or multi-disease co-occurrence [8, 9].

Vision-Language Pre-training (VLP) frameworks, most notably CLIP [10], address this limitation by aligning images with clinical text in a shared semantic space, enabling richer and more transferable feature representations. Building on this paradigm, a series of fundus-specific VLP models have been proposed. FLAIR [6] aggregates expert-curated text from clinical literature and community standards to achieve strong zero-shot performance, yet it predominantly maps categorical labels to text templates, failing to capture complex pathological reasoning. MM-Retinal [7] and RET-CLIP [11] incorporate professional fundus atlas knowledge and

[†]Lei Shi and Wenbin Zhai have contributed equally to this work.

*Correspondence: fishstonehfu1006@ahthcm.edu.cn

³ School of Medical Information Technology, Anhui University of Chinese Medicine, Hefei 230012, Anhui, China

Full list of author information is available at the end of the article

real diagnostic reports, respectively, advancing representation quality in complementary ways. Nevertheless, all of these models share a critical omission: they uniformly neglect the individualized patient context embedded in Electronic Health Records (EHR), including chief complaints, history of present illness, systemic indicators, and past medical history. This gap is consequential, because the clinical reasoning of expert ophthalmologists is inherently holistic rather than image-centric.

In real-world clinical practice, a senior ophthalmologist does not interpret a fundus image in isolation; rather, they synthesize visual findings with the patient's systemic profile, encompassing the duration and glycemic control of diabetes, blood pressure stability, and presenting symptoms such as blurred vision. A representative example illustrates this limitation: subtle microaneurysms that a vision-only model may dismiss as noise carry decisive diagnostic weight when contextualized against a documented history of chronic hyperglycemia. By forgoing such individualized clinical context, existing VLP frameworks remain ill-equipped to resolve visual ambiguities, limiting their logical interpretability and practical utility in clinical decision support.

To bridge this gap, we present a tightly integrated dataset and model framework. We first construct **MedRecordFundus**, a large-scale multimodal dataset that systematically pairs high-quality CFP images with de-identified EHR narratives, professionally annotated across four structured clinical dimensions: Diagnosis, Chief Complaint, History of Present Illness, and Past Medical History. Leveraging this resource, we propose **MedRecord-CLIP**, a knowledge-enhanced foundation model built upon a ViT-B/16 visual encoder and a RoBERTa text encoder, augmented by a novel **diagnosis-guided cross-attention mechanism** that adaptively retrieves diagnostically relevant evidence from individualized patient histories. To further address the pervasive *false-negative* problem in medical contrastive learning, where semantically similar image-text pairs from different patients are incorrectly penalized as negatives, we introduce a symmetric **Multi-Positive InfoNCE objective** that extends the traditional one-hot contrastive target to a multi-positive formulation. Extensive experiments across four public benchmarks demonstrate that our approach yields statistically significant improvements over state-of-the-art baselines, including RETFound [2] and FLAIR [6], on the multi-disease benchmarks (RFMiD and ODIR), and further leads on the IDRiD grading benchmark while performing on par with the strongest baseline on APTOS 2019, with ablation studies validating

the independent contribution of each proposed component. Our main contributions are as follows:

- We construct **MedRecordFundus**, a large-scale multimodal fundus dataset comprising 21,290 CFP images and 1,043 expert-curated image-text pairs, each verified and multi-dimensionally annotated by senior ophthalmologists across four structured clinical dimensions. To our knowledge, this is the first fundus dataset to systematically incorporate individualized patient EHR narratives spanning diagnosis, chief complaints, present illness history, and past medical history as supervisory signals for vision-language pre-training.
- We propose **MedRecord-CLIP**, a foundation model featuring a **diagnosis-guided cross-attention mechanism** that uses the diagnostic representation as a query over the history representation, enabling adaptive retrieval of salient clinical evidence from lengthy patient records. To mitigate the false-negative problem arising from semantic overlap in medical narratives, we further incorporate a symmetric **Multi-Positive InfoNCE objective** that extends the standard one-hot contrastive target to a multi-positive formulation, yielding more robust and discriminative joint embeddings.
- We conduct extensive experiments across **four publicly available benchmarks**, namely RFMiD, ODIR, APTOS 2019, and IDRiD [12–15]. Results demonstrate that the multimodal synergy between fundus imagery and personalized clinical context yields statistically significant gains on the multi-disease benchmarks (RFMiD and ODIR), leads on the IDRiD grading benchmark, and remains on par with the strongest baseline on APTOS 2019. Comprehensive ablation studies further validate the individual contribution of clinical history integration, cross-attention fusion, and the multi-positive contrastive objective.

Related works

Visual foundation models for fundus imaging

With the success of foundation models in computer vision, researchers have begun to explore self-supervised pre-training on large-scale fundus image datasets. RETFound [2] is a seminal work in this direction, employing a Masked Image Modeling (MIM) architecture pre-trained on over 1.6 million unannotated fundus images. By learning to reconstruct masked image patches, RETFound extracts highly generalizable spatial representations of retinal structures, providing robust visual feature support

for a wide range of downstream tasks, including diabetic retinopathy grading, glaucoma screening, and systemic disease prediction. Beyond fundus-specific models, general-purpose vision foundation models such as the Vision Transformer (ViT) [16] and its variants have also been adapted to medical imaging, demonstrating that large-scale pre-training significantly reduces the dependency on task-specific labeled data. However, these vision-only approaches are inherently limited: by relying solely on pixel-level reconstruction objectives, they discard the semantic and clinical knowledge embedded in medical language, constraining their interpretability and transferability across heterogeneous clinical environments.

Vision-language pre-training in medical imaging

Vision-Language Pre-training (VLP) addresses this limitation by introducing textual semantic supervision, jointly aligning visual and linguistic representations to produce more interpretable and transferable features. Inspired by the success of CLIP [10] in the general domain, a growing body of work has adapted contrastive vision-language learning to medical imaging. MedCLIP [5] is a pioneering effort in this area, decoupling the strict image-text pairing constraint of standard contrastive learning and replacing it with a semantic matching loss grounded in medical knowledge graphs. This design enables effective cross-modal representation learning even when image-text correspondences are weak or incomplete, substantially reducing the annotation burden. BioViL [17] further explores phrase-grounded pre-training on radiology report-image pairs, demonstrating that fine-grained local alignment between clinical phrases and image regions yields superior performance in pathology localization. More recently, LLaVA-Med [18] extends large multimodal language models to the biomedical domain through instruction-tuning on biomedical image-text pairs, enabling open-ended visual question answering. Despite these advances, most existing VLP frameworks in medical imaging rely on population-level text supervision derived from standardized reports or curated literature, failing to incorporate the individualized patient context present in real-world Electronic Health Records (EHR).

Multimodal foundation models for retinal analysis

In recent years, a series of multimodal foundation models that deeply integrate domain knowledge have been proposed specifically for fundus image analysis, progressively enriching the granularity of textual supervision:

- **FLAIR** [6] focuses on transforming expert knowledge into structured descriptive text. By pooling 37 open-access fundus datasets spanning 96 pathologi-

cal categories and using professional text extracted from clinical literature and community diagnostic standards as supervisory signals, FLAIR achieves a deep and comprehensive capture of diverse fundus lesion features. The model demonstrates strong zero-shot transferability across unseen disease categories, reflecting high consistency with professional clinical knowledge. Nevertheless, its textual supervision is largely confined to standardized categorical descriptions, limiting its capacity to capture complex inter-disease reasoning or patient-specific pathological context.

- **RET-CLIP** [11] pioneers the use of real-world clinical diagnostic reports as pre-training supervision for retinal foundation models. The model adopts a hierarchical three-level optimization strategy encompassing the left eye, right eye, and patient levels, trained on data from over 190,000 patients. By jointly considering monocular structural details and patient-level clinical relevance, RET-CLIP produces feature representations that more closely mirror the diagnostic logic employed in real outpatient settings. However, its supervision remains focused on structured diagnostic conclusions, and it does not incorporate broader patient history information such as chief complaints, systemic disease background, or longitudinal clinical trajectories.
- **MM-retinal** [7] contributes a high-quality knowledge-enhanced multimodal dataset whose sources are primarily drawn from professional fundus atlas books annotated by domain experts. Building on this resource, the KeepFIT model employs image similarity-guided text revision and a hybrid training strategy to implant textbook-level expert knowledge directly into the visual encoder, substantially improving generalization performance across heterogeneous data sources and imaging devices. Despite its strong domain knowledge infusion, the atlas-derived text is inherently population-level and standardized, offering no mechanism to account for the individualized clinical context of specific patients.

Positioning of this work

Compared with the most relevant fundus VLP frameworks along the axis of textual supervision, our work differs in three complementary aspects. FLAIR maps categorical labels to text templates distilled from clinical literature; RET-CLIP advances to real diagnostic reports; and MM-Retinal injects professional fundus-atlas knowledge. Critically, FLAIR and MM-Retinal rely on population-level text that is invariant across patients sharing the same label, while RET-CLIP,

although it leverages real diagnostic reports, captures only the diagnostic conclusion and not the systemic history or individualized context of a specific patient. In contrast, at the data level, MedRecordFundus is, to our knowledge, the first fundus dataset to pair images with patient-specific EHR narratives spanning diagnosis, chief complaint, history of present illness, and past medical history. Crucially, because each record documents the systemic and ocular comorbidities of an individual patient, such as hypertension, prior cataract or glaucoma, and concurrent retinal findings, a single image becomes associated with the semantics of multiple co-occurring conditions rather than one isolated label; this cross-label supervision is precisely the structure that downstream multi-disease recognition requires, and it explains why even a modest number of richly annotated records contributes an influence disproportionate to their share of the corpus. At the architecture level, rather than aligning images with a single static caption, we introduce a diagnosis-guided cross-attention mechanism that uses the diagnostic representation as a query to adaptively retrieve diagnostically relevant evidence from the lengthy, heterogeneous history narrative. At the objective level, we replace the standard one-hot contrastive target with a symmetric Multi-Positive InfoNCE loss that explicitly

addresses the false-negative problem caused by the semantic overlap pervasive in clinical text. Together, these designs position MedRecord-CLIP as a framework that learns from holistic, patient-level clinical reasoning rather than from standardized labels alone.

Method

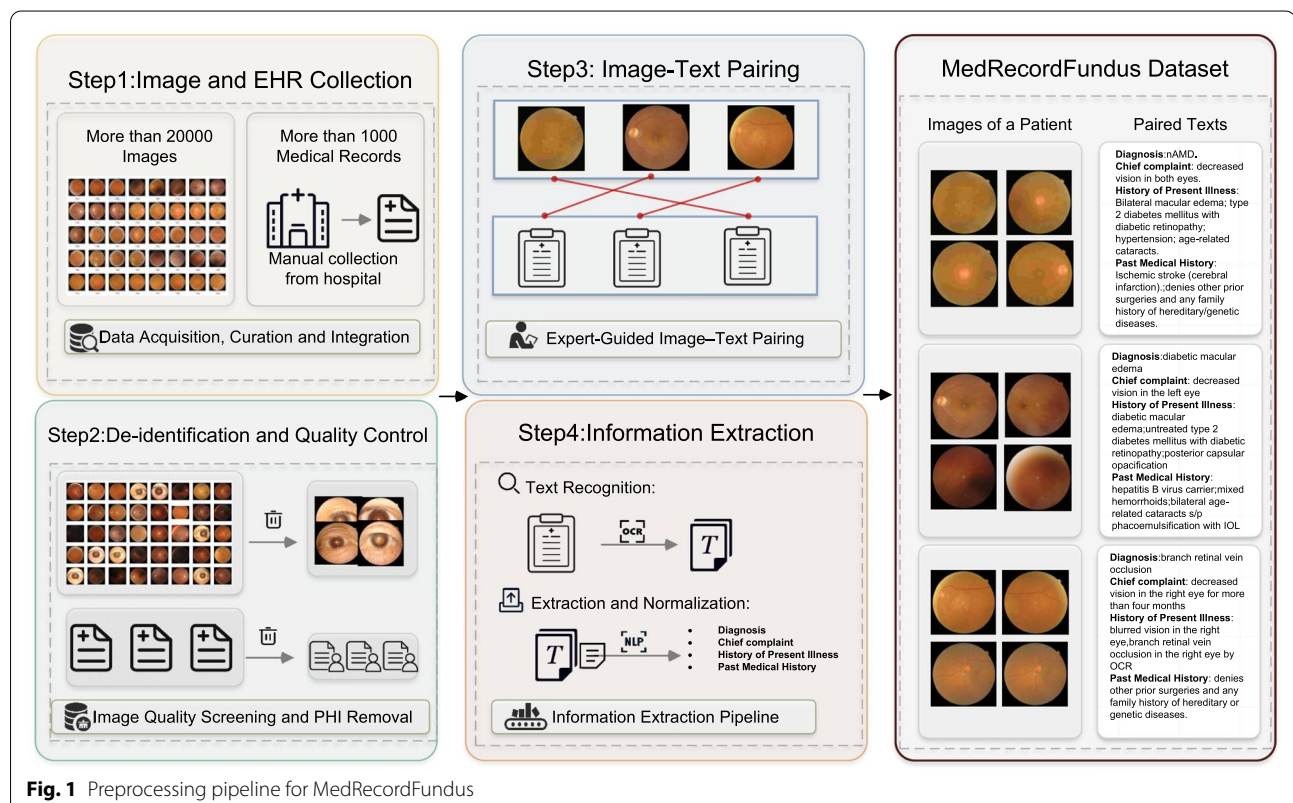
In this section, we present the complete framework of **MedRecord-CLIP**, which consists of three key components. First, we describe the construction pipeline of the **MedRecordFundus** dataset. Second, we introduce the architecture of **MedRecord-CLIP**, featuring a dual-encoder design augmented by a novel **diagnosis-guided cross-attention** mechanism for adaptive multimodal fusion. Finally, we formulate a **symmetric Multi-Positive InfoNCE** training objective to align visual and textual modalities robustly under the challenge of semantic overlap prevalent in clinical narratives.

Dataset construction

Data collection

As illustrated in Fig. 1, we established a rigorous four-stage pipeline to construct the multimodal **MedRecordFundus** dataset.

Step 1: Image and EHR collection. We collected over 20,000 CFP images and 1,000 EHRs from a tertiary



teaching hospital in Anhui Province, China, under institutional ethics approval. All data were acquired using standardized fundus cameras to ensure imaging consistency across patients.

Step 2: De-identification and quality control. To ensure clinical reliability and protect patient privacy, each image underwent a systematic quality screening process to filter out sub-optimal acquisitions affected by poor illumination, motion blur, or insufficient field-of-view coverage. Concurrently, all protected health information (PHI) was removed from EHR records in strict compliance with applicable data privacy regulations.

Step 3: Image-text pairing. To bridge the semantic gap between visual observations and clinical documentation, senior ophthalmologists manually matched each CFP image to its corresponding EHR narrative, yielding 1,043 image-text pairs. Each pair was independently verified to ensure diagnostic coherence between the fundus findings and the associated clinical record.

Step 4: Automated clinical structuring. To enable structured multimodal supervision, we developed an automated clinical structuring engine that leverages high-precision Optical Character Recognition (OCR) for narrative digitization and specialized Natural Language Processing (NLP) models for information extraction. This engine parses each unstructured EHR narrative into four clinically meaningful dimensions: **Diagnosis, Chief Complaint, History of Present Illness, and Past Medical History**. This taxonomic decomposition ensures that the model receives supervision signals covering both acute pathological manifestations and long-term systemic medical context.

Data preprocessing

Image Preprocessing. To mitigate resolution heterogeneity introduced by different acquisition devices, all CFP images were uniformly resized to 512×512 pixels as a base resolution. During training, a stochastic online augmentation pipeline was applied: sub-regions were randomly cropped and resized to 224×224 pixels to simulate varied visual fields and framing offsets. To account for the bilateral symmetry of retinal anatomy, random horizontal flipping was applied. Photometric robustness against sensor-induced variations was further enhanced through mild color jittering, including random perturbations of brightness, contrast, and saturation. Final pixel values were normalized using the channel-wise mean and standard deviation consistent with the pre-trained visual backbone, ensuring input distribution compatibility. **Text Preprocessing.** Clinical narratives digitized via OCR and structured through the NLP-based information extraction module underwent rigorous normalization to

align raw text with the language capabilities of the text encoder. Specifically, this process involved: (i) rectifying typographical and OCR-induced errors; (ii) standardizing non-conventional punctuation; (iii) expanding non-standard abbreviations into formal medical nomenclature; and (iv) unifying code-switched Chinese-English expressions into standard Chinese. Following manual spot checks by clinical experts to verify semantic coherence, we maintained a strictly conservative preprocessing policy that excluded any additional clinical inferences or text rewrites. This ensures that the original clinical conclusions and semantic nuances of each EHR record remain intact, preserving the high-fidelity authenticity and clinical validity of **MedRecordFundus**.

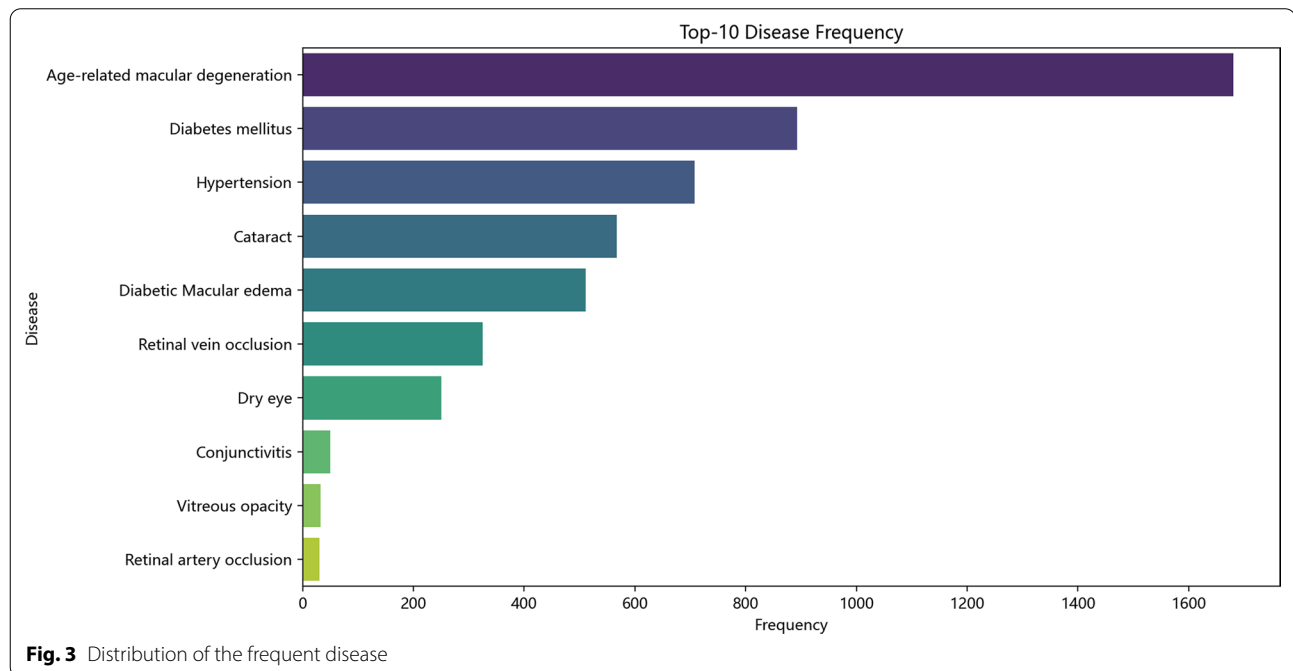
Dataset statistics

To intuitively demonstrate the semantic richness and pathological diversity of the MedRecordFundus dataset, we conducted the following statistical analyses:

- **Lexical Density:** As shown in Fig. 2, the word cloud illustrates the high-frequency clinical terms appearing in the medical records. In addition to pathological descriptions related to the retina, it also extensively covers systemic indicators and patient chief complaints that are crucial for diagnostic decision-making.
- **Pathological Distribution:** As shown in Fig. 3, we have counted and listed the 10 most frequent diseases in the dataset. This distribution indicates that the dataset covers mainstream blinding eye diseases, ranging from diabetic retinopathy and glaucoma to age-related macular degeneration, ensuring the robustness and representativeness of the model when handling multi-task clinical scenarios.

Model architecture

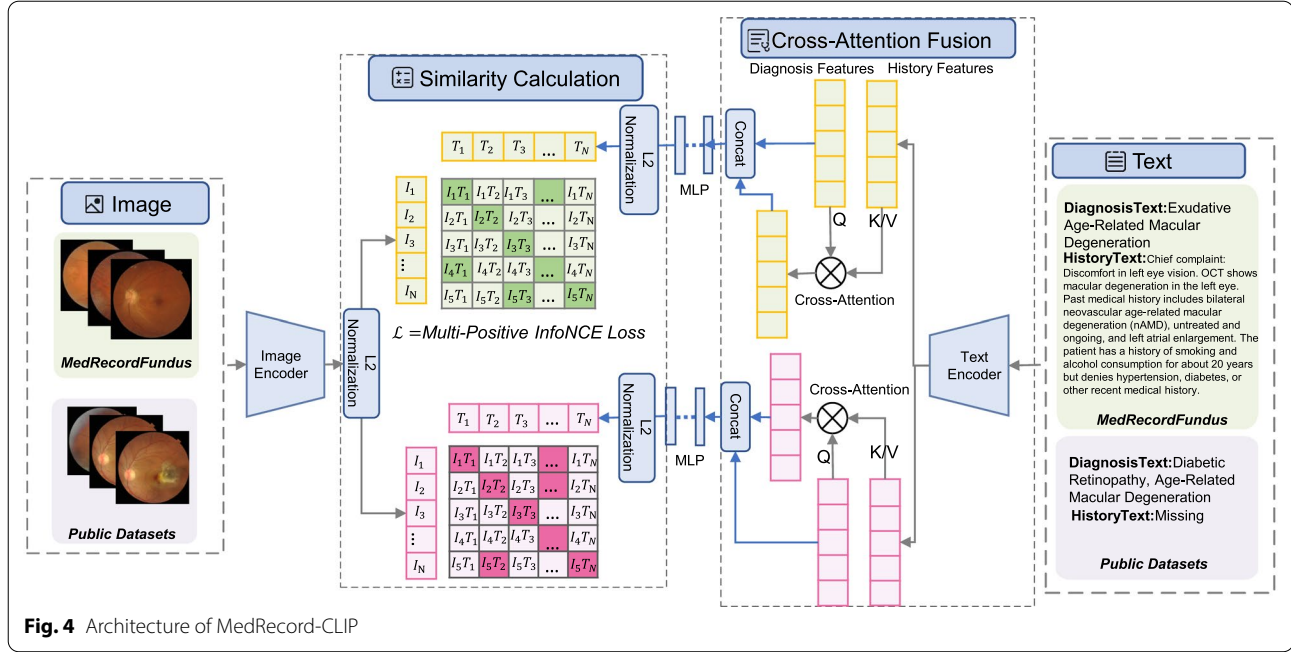
As illustrated in Fig. 4, **MedRecord-CLIP** is a vision-language foundation model pre-trained on both publicly available fundus datasets and **MedRecordFundus**. For public datasets that provide only discrete category labels, we textualize each class by mapping it to a diagnostic sentence template; the corresponding history text field is left empty or replaced with a placeholder token, ensuring architectural consistency during mixed-source pre-training. All pre-training is conducted in Chinese, as the EHR narratives in **MedRecordFundus** originate from a Chinese tertiary hospital; the standardized diagnostic templates derived from the English labels of public datasets are likewise translated into Chinese so that all textual



supervision shares a single language space, consistent with the preprocessing step that unifies code-switched expressions into standard Chinese. The textual examples shown elsewhere in this paper are English translations of the original Chinese records, provided solely for readability; the model itself operates on the original Chinese text. The model comprises a ViT-B/16 vision encoder and two RoBERTa-based text encoders, which extract

representations from fundus images and clinical narratives, respectively, and project them into a shared contrastive embedding space of dimension d .

Given a batch of size N , each sample is denoted as (I_i, T_i^d, T_i^h) , where I_i is the fundus image, T_i^d is the diagnosis text, and T_i^h is the patient history text. Let d denote the contrastive embedding dimension and H the hidden size of the text encoders.



Visual Encoder: The visual encoder is based on the Vision Transformer (ViT). ViT partitions an image into fixed-size patches, treats each patch as a token, and applies a multi-layer Transformer to extract representations, enabling the model to capture global structure and local details effectively. In practice, we initialize with a pretrained ViT-B/16 and fine-tune it for fundus imaging. The visual encoder Φ_v outputs a d -dimensional image feature v_i :

$$v_i = \Phi_v(I_i) \in \mathbb{R}^{1 \times d}. \quad (1)$$

Text Encoder: The diagnosis text T_i^d and the history text T_i^h are processed by two *separate* RoBERTa-based text encoders Φ_d and Φ_h , respectively, both initialized from Chinese-RoBERTa-wwm-ext. The two encoders play asymmetric roles in the subsequent fusion: the diagnosis is summarized into a single pooled [CLS] vector that serves as the query, whereas the history is kept as a full token sequence that serves as the key/value attended to by that query. We keep the two encoders separate so that each gradient flows only to its own parameters, allowing Φ_d to specialize in producing strong retrieval queries and Φ_h in preserving fine-grained, position-sensitive token detail; a shared encoder would instead force the same parameters to serve both conflicting goals. We verify this design choice with a parameter-sharing ablation (Model D) in Sect. [Ablation study](#). Concretely, the diagnostic feature $t_i^d \in \mathbb{R}^{1 \times d}$ is obtained by linearly projecting the pooled [CLS] token vector of dimension H . Crucially,

to preserve the detailed context of the lengthy patient history narrative, we extract the sequence of all token embeddings for the history text, denoted as a matrix $T_i^h \in \mathbb{R}^{L \times d}$, where L is the sequence length:

$$\begin{aligned} t_i^d &= \Phi_d(T_i^d)[\text{CLS}] W_d \in \mathbb{R}^{1 \times d}, \\ T_i^h &= \Phi_h(T_i^h) W_h \in \mathbb{R}^{L \times d}. \end{aligned} \quad (2)$$

Diagnosis-Guided Cross-Attention Fusion: Patient history text is typically a lengthy narrative containing substantial information that may be tangential to the current diagnosis. To selectively attend to clinically relevant evidence, we introduce a diagnosis-guided cross-attention mechanism in which the diagnostic representation acts as the query to retrieve salient components from the history token sequence. Formally, let $Q, K, V \in \mathbb{R}^{d \times d_h}$ be linear projection matrices mapping to the single-head attention dimension d_h . The query vector, key matrix, and value matrix are computed as:

$$\begin{aligned} q_i &= t_i^d Q \in \mathbb{R}^{1 \times d_h}, & K_i &= T_i^h K \in \mathbb{R}^{L \times d_h}, \\ V_i &= T_i^h V \in \mathbb{R}^{L \times d_h}. \end{aligned} \quad (3)$$

The attention weights α_i and the resulting context vector $c_i \in \mathbb{R}^{1 \times d_h}$ are computed via scaled dot-product attention over the sequence:

$$\alpha_i = \text{softmax}\left(\frac{q_i K_i^\top}{\sqrt{d_h}}\right) \in \mathbb{R}^{1 \times L}, \quad c_i = \alpha_i V_i. \quad (4)$$

To map the context vector back to the shared dimension d , we apply an output projection matrix $W_o \in \mathbb{R}^{d_h \times d}$. A residual connection followed by layer normalization (LN) and a two-layer feed-forward network (FFN) produces the diagnosis-conditioned history representation r_i :

$$r_i = \text{FFN}\left(\text{LN}\left(c_i W_o + t_i^d\right)\right). \quad (5)$$

The diagnostic representation t_i^d and the refined history representation r_i are then concatenated and passed through a two-layer MLP with first-layer weight $W_1 \in \mathbb{R}^{2d \times d}$, second-layer weight $W_2 \in \mathbb{R}^{d \times d}$, ReLU activation σ , and inter-layer dropout, to produce the final fused textual representation:

$$t_i^{\text{fused}} = \sigma\left(\left[t_i^d; r_i\right] W_1\right) W_2. \quad (6)$$

Normalization and Similarity: We perform ℓ_2 normalization on the image and fused text features. Let the learnable scaling parameter be $s = \exp(\alpha)$ (where α is `logit_scale`); then the score matrix is defined as:

$$\hat{v}_i = \frac{v_i}{\|v_i\|_2}, \quad \hat{t}_i = \frac{t_i^{\text{fused}}}{\|t_i^{\text{fused}}\|_2}, \quad L_{ij} = s \langle \hat{v}_i, \hat{t}_j \rangle. \quad (7)$$

Training Objective: From statistical analysis and empirical observations, we find that within the same batch many captions for different CFP images are highly similar. If all unpaired image–text pairs are uniformly treated as negatives, this can introduce false negatives and degrade alignment quality. To address this, we adopt Multi-Positive InfoNCE, extending the traditional one-hot target (where the positive set is strictly $\{i\}$) to multi-positive target sets, denoted as P_i^{txt} and P_i^{img} .

Positive Set Construction: We dynamically construct these multi-positive sets by quantifying the clinical semantic overlap within a batch. Let $\hat{T} \in \mathbb{R}^{N \times d}$ denote the ℓ_2 -normalized text feature matrix, where the i -th row corresponds to $\hat{t}_i$. The in-batch text semantic similarity matrix is computed as:

$$S^{\text{txt}} = \hat{T} \hat{T}^\top. \quad (8)$$

For each anchor i , we construct the positive sets based on its text semantic neighborhood. The text positive set for an image anchor i is defined as:

$$P_i^{\text{txt}} = \{i\} \cup \text{TopK}\left(\{j \neq i \mid S_{ij}^{\text{txt}} \geq \tau_t\}, k_t\right). \quad (9)$$

Crucially, to prevent false positive matches caused by visual ambiguities, the image positive set for a text anchor i is not constructed based on image-to-image similarity. Instead, it reuses the aforementioned text semantic

neighborhood by selecting the text indices most semantically similar to text anchor i , and explicitly mapping them to their corresponding paired image indices via a bijective mapping function $\pi(\cdot)$:

$$P_i^{\text{img}} = \{i\} \cup \left\{ \pi(j) \mid j \in \text{TopK}\left(\{j' \neq i \mid S_{ij'}^{\text{txt}} \geq \tau_v\}, k_v\right) \right\}. \quad (10)$$

Here, τ_t and τ_v are predefined similarity thresholds, while k_t and k_v are the upper bounds of additional positive samples retained for each anchor. The selection inherently excludes the anchor itself ($j \neq i$). If the number of candidates satisfying the threshold is less than k , all such candidates are retained. This design ensures that the expansion of positive samples in both directions is strictly guided by the EHR text semantics, thereby mitigating the risk of mismatching induced by the visual ambiguities inherent in medical images. The symmetric multi-positive InfoNCE is written as:

$$\begin{aligned} \mathcal{L}_{\text{img}}^{\text{MP}} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{\sum_{j \in P_i^{\text{txt}}} \exp(L_{ij})}{\sum_{j=1}^N \exp(L_{ij})}, \\ \mathcal{L}_{\text{txt}}^{\text{MP}} &= -\frac{1}{N} \sum_{i=1}^N \log \frac{\sum_{j \in P_i^{\text{img}}} \exp(L_{ji})}{\sum_{j=1}^N \exp(L_{ji})} \end{aligned} \quad (11)$$

The final loss is given by:

$$\mathcal{L} = \frac{1}{2} (\mathcal{L}_{\text{img}}^{\text{MP}} + \mathcal{L}_{\text{txt}}^{\text{MP}}). \quad (12)$$

Experiments and results

Experimental setup

Pre-training datasets

MedRecord-CLIP is pre-trained on a diverse, large-scale collection of two complementary sources:

- (1) **MedRecordFundus.** Our self-constructed multi-modal dataset comprising 21,290 CFP images and 1,043 expert-curated image-text pairs, each annotated across four structured clinical dimensions by senior ophthalmologists. This dataset provides the individualized patient-level EHR supervision that distinguishes our framework from prior VLP approaches.
- (2) **FLAIR** [6]. A comprehensive publicly available foundation dataset that, in its original form, aggregates 37 open-access fundus image collections across 96 disease categories, totaling 284,660 images. For samples in this dataset that provide only discrete category labels, we textualize each label by mapping it to a standardized diagnostic sentence template, with the history text field replaced by a place-

holder token to maintain architectural consistency. FLAIR is organized as a merger of its constituent open-access datasets, each kept as a separate source collection, and the four evaluation benchmarks (RFMiD, ODIR, APTOS 2019, IDRiD) are themselves among these constituent datasets. We therefore keep pre-training and evaluation strictly disjoint simply by deleting these four source datasets in their entirety from the pre-training corpus. The 284,660-image figure above refers to FLAIR in its original form; after this deletion, only the remaining source datasets are used for pre-training. We apply the identical exclusion to every FLAIR-based baseline (FLAIR, CN-CLIP).

Evaluation datasets

To comprehensively assess downstream diagnostic performance, we evaluate on four widely recognized public benchmarks spanning different disease categories, imaging conditions, and annotation granularities.

- (1) **RFMiD** [12]. A multi-label benchmark comprising 3,200 high-quality fundus images annotated across 28 referable fundus conditions. Following the standard evaluation protocol, we report both disease screening performance (DRS AUC) and multi-disease classification performance (mAUROC), making it the primary benchmark for evaluating the model's ability to handle complex multi-pathology scenarios.
- (2) **ODIR** [13]. A large-scale clinical dataset containing 10,000 fundus images from 5,000 patients, with paired left and right eye images per patient. It provides comprehensive multi-label annotations covering eight clinical categories: normal retina, diabetic retinopathy, glaucoma, cataract, age-related macular degeneration, hypertension, myopia, and other abnormalities. This dataset evaluates the model's capacity for patient-level multi-disease recognition under realistic clinical data distribution.
- (3) **APTOS 2019** [14]. A diabetic retinopathy grading benchmark comprising 3,662 fundus images annotated with a five-level severity scale: no DR, mild, moderate, severe, and proliferative DR. This benchmark assesses the model's fine-grained diagnostic capability across continuous disease progression stages, reflecting the clinical need for accurate severity stratification.
- (4) **IDRiD** [15]. A specialized high-resolution CFP dataset consisting of 516 images focused on diabetic retinopathy, providing expert-level annotations for the same five severity levels as APTOS 2019 as well

as pixel-level lesion segmentation labels. Its small scale and high annotation precision make it a critical benchmark for evaluating the model's performance under limited data conditions and precise clinical assessment requirements.

Baselines and evaluation metrics

Baselines

To rigorously benchmark **MedRecord-CLIP**, we compare against four representative models spanning supervised, self-supervised, and vision-language pre-training paradigms:

- **CN-CLIP** [19]: A large-scale Chinese CLIP model pre-trained on massive non-medical image-text pairs. It serves as a general-domain vision-language baseline to quantify the benefit of domain-specific medical pre-training.
- **RETFund** [2]: A large-scale visual foundation model for fundus images based on the Masked Image Modeling (MIM) architecture, pre-trained on over 1.6 million unannotated fundus images. It represents the state-of-the-art in vision-only self-supervised fundus pre-training.
- **FLAIR** [6]: A fundus-specific vision-language foundation model that incorporates expert-curated clinical text as supervisory signals. It represents the most competitive VLP baseline in our evaluation, using standardized categorical text without individualized patient context.
- **EfficientNet** [20]: A standard CNN baseline trained in a fully supervised manner on each benchmark's training split. It establishes the performance reference for task-specific supervised learning without any pre-training benefit.

Evaluation metrics

We adopt two complementary threshold-free metrics to provide a comprehensive assessment of diagnostic accuracy across both screening and multi-disease recognition tasks:

- **Disease Risk Score AUC (DRS AUC)** [12]: This metric evaluates the model's overall binary screening capability. All pathological labels are collapsed into a single *abnormal* class, and the AUC of the ROC curve is computed for the binary discrimination between *normal* and *abnormal* fundus images. A higher DRS AUC reflects stronger sensitivity to the presence of any retinal pathology, which is critical for large-scale population screening.

- **Mean AUROC (mAUROC) [21]:** For the K -label multi-disease recognition task, we compute the Area Under the ROC Curve independently for each of the K disease categories and report the macro-average. This metric provides a balanced assessment of the model's discriminative proficiency across all disease categories, regardless of class imbalance, and directly reflects clinical utility in differential diagnosis scenarios.

Comparison study

We report the downstream performance of MedRecord-CLIP under a unified, task-specific adaptation framework. Specifically, all pre-trained models—including our MedRecord-CLIP and every baseline—are adapted to each benchmark through full fine-tuning: starting from their respective pre-trained weights, the entire network is unfrozen and trained jointly with a linear classification head appended to the image encoder.

To assess the statistical reliability of our results, all experiments were conducted across 5 random seeds, and all metrics are reported as mean $\pm$ standard deviation. For each benchmark, we pair the per-seed scores of MedRecord-CLIP with those of the strongest baseline and apply a two-sided paired t -test over the 5 seeds. The improvements on the multi-disease benchmarks RFMiD and ODIR are statistically significant at $p < 0.05$, while on the diabetic-retinopathy grading benchmarks MedRecord-CLIP remains competitive with the strongest baseline.

Fine-tuning configuration. Every model and benchmark adopts an identical adaptation protocol. We optimize with AdamW under a cosine-decay schedule with linear warm-up, and adopt discriminative layer-wise learning-rate decay (LLRD): the linear head uses the largest learning rate while successively lower transformer blocks are decayed by a fixed factor, so that the entire network is unfrozen but low-level features are perturbed less than task-specific upper layers. We train with batch size 32 at 224×224 resolution and apply early stopping on the validation mAUROC (up to 100 epochs maximum). To counter overfitting on the small downstream sets, we further apply stochastic depth, dropout before the head, and a Mixup/CutMix policy on top of fundus-appropriate geometric and mild photometric augmentations. **Task-adaptive objectives.** For the multi-label benchmarks (RFMiD, ODIR) we use an asymmetric loss (ASL) over per-label sigmoids to mitigate the long-tailed class distribution; for the five-level diabetic-retinopathy grading benchmarks (APTOS 2019, IDRiD) we adopt an ordinal-regression objective (CORN) with label smoothing, which exploits the ordering of severity levels. **Data**

Table 1 Fine-tuning performance comparison on multi-disease fundus datasets (RFMiD and ODIR)

Models	RFMiD		ODIR	
	DRS AUC	mAUROC	DRS AUC	mAUROC
CN-CLIP	0.969 $\pm$ 0.008	0.896 $\pm$ 0.010	0.864 $\pm$ 0.009	0.856 $\pm$ 0.011
RETFound	0.966 $\pm$ 0.007	0.891 $\pm$ 0.009	0.860 $\pm$ 0.008	0.844 $\pm$ 0.010
FLAIR	0.901 $\pm$ 0.010	0.840 $\pm$ 0.012	0.863 $\pm$ 0.009	0.830 $\pm$ 0.011
Efficient-Net	0.967 $\pm$ 0.009	0.800 $\pm$ 0.013	0.817 $\pm$ 0.011	0.809 $\pm$ 0.012
OURS	0.978 $\pm$ 0.006	0.910 $\pm$ 0.008	0.903 $\pm$ 0.007	0.888 $\pm$ 0.009

The results of our method are shown in bold

Table 2 Fine-tuning performance comparison on diabetic retinopathy grading datasets (APTOS 2019 and IDRiD)

Models	APTOS 2019		IDRiD	
	DRS AUC	mAUROC	DRS AUC	mAUROC
CN-CLIP	0.878 $\pm$ 0.008	0.819 $\pm$ 0.010	0.839 $\pm$ 0.019	0.797 $\pm$ 0.024
RETFound	0.940 $\pm$ 0.006	0.920 $\pm$ 0.008	0.867 $\pm$ 0.016	0.849 $\pm$ 0.020
FLAIR	0.905 $\pm$ 0.008	0.882 $\pm$ 0.009	0.828 $\pm$ 0.020	0.768 $\pm$ 0.026
Efficient-Net	0.852 $\pm$ 0.010	0.827 $\pm$ 0.012	0.806 $\pm$ 0.022	0.749 $\pm$ 0.028
OURS	0.938 $\pm$ 0.007	0.921 $\pm$ 0.008	0.892 $\pm$ 0.015	0.854 $\pm$ 0.019

The results of our method are shown in bold

splits. We use the official training/validation/test splits where available; for ODIR we adopt a patient-level split so that both eyes of a patient never cross folds, and for the small IDRiD set we report 5-fold cross-validation. For each configuration the checkpoint with the best validation mAUROC is selected for testing. The complete hyper-parameters, configuration files, and training code are made publicly available (see the Reproducibility statement).

Multi-disease diagnostic performance

Table 1 summarizes the classification performance on the multi-disease benchmarks (**RFMiD** and **ODIR**), with all entries reported as mean $\pm$ standard deviation over 5 random seeds. MedRecord-CLIP achieves a fine-tuned DRS AUC of 0.978 ± 0.006 and an mAUROC of 0.910 ± 0.008 on RFMiD, outperforming the strongest vision-language foundation model FLAIR by 7.7 and 7.0 percentage points, respectively. Similarly, on the patient-level ODIR dataset, our model yields a DRS AUC of 0.903 ± 0.007 and an mAUROC of 0.888 ± 0.009 , establishing a clear performance margin over both vision-only and text-supervised baselines.

Diabetic retinopathy grading performance

For the continuous diabetic retinopathy severity grading tasks, Table 2 reports the comparative outcomes on the APTOS 2019 and IDRiD benchmarks, with all entries reported as mean $\pm$ standard deviation over 5 random seeds. MedRecord-CLIP maintains its consistent representation edge, achieving mAUROC values of 0.921 ± 0.008 and 0.854 ± 0.019 , respectively.

Ablation study

To quantify the individual contribution of each core component in MedRecord-CLIP, we conducted a systematic ablation study on the RFMiD and ODIR benchmarks. Three variants are constructed by removing one component at a time from the full model: (A) removal of patient history text T^h , (B) replacement of cross-attention fusion with simple feature concatenation, and (C) replacement of the Multi-Positive InfoNCE loss with the standard CLIP contrastive loss. To further analyze the text-encoder design and the source of the observed gains, we evaluate two additional settings: (D) sharing a single set of parameters between the diagnosis and history text encoders; and (E) randomly shuffling the EHR history records across patients during pre-training. The component ablation (A–C) is summarized in Table 3, and the control and robustness study (D–E) in Table 4. All ablation variants are trained under the same 5 random seeds as the full model, and all entries are reported as mAUROC (mean $\pm$ standard deviation). Each variant is compared against the full model with a two-sided paired t -test (paired by seed). Removing the patient history (Model A) and shuffling the EHR records (Model E) produce statistically significant degradations on both RFMiD and ODIR ($p < 0.05$), as do the cross-attention (Model B) and dual-encoder (Model D) ablations; the smaller margin from the Multi-Positive objective (Model C) is significant on ODIR and reaches only trend-level significance on RFMiD.

Effectiveness of clinical history integration

Removing the patient history text (Model A) leads to a substantial drop in mAUROC of 3.3 and 4.1 percentage points on RFMiD and ODIR, respectively. This result

Table 4 Control and robustness study of MedRecord-CLIP on RFMiD and ODIR datasets

Model	Text encoder	EHR	RFMiD	ODIR
Full model	Dual	Real	0.910 ± 0.008	0.888 ± 0.009
(D) shared encoder	Shared	Real	0.899 ± 0.009	0.870 ± 0.011
(E) shuffled EHR	Dual	Shuffled	0.864 ± 0.012	0.835 ± 0.013

Results are reported as mAUROC, bold values indicate the results of the full model

confirms that individualized clinical context, encompassing chief complaints, systemic indicators, and past medical history, provides supervisory signals that fundus images alone cannot encode. The performance degradation is particularly pronounced on ODIR, which features multi-disease annotations spanning systemic conditions such as hypertension and diabetes, underscoring the value of EHR integration for conditions with strong systemic correlates.

Impact of diagnosis-guided cross-attention fusion

Replacing the cross-attention mechanism with naive feature concatenation (Model B) results in performance drops of 1.7 and 2.4 percentage points in mAUROC on RFMiD and ODIR, respectively. This demonstrates that simply combining diagnostic and history representations without selective attention introduces noise from clinically irrelevant content in lengthy EHR narratives. The cross-attention mechanism, by using the diagnostic representation as a query over the history representation, effectively filters out tangential information and concentrates on history segments most relevant to the current diagnosis, thereby improving the discriminative capacity of the fused textual feature.

Contribution of multi-positive InfoNCE objective

Substituting the Multi-Positive InfoNCE with the standard CLIP contrastive loss (Model C) reduces mAUROC by 1.0 and 1.6 percentage points on RFMiD and ODIR, respectively. This validates our hypothesis that, in

Table 3 Ablation study of MedRecord-CLIP on RFMiD and ODIR datasets

Model	History (T^h)	Fusion	Loss	RFMiD	ODIR
Full Model	✓	Cross-Attn	$\mathcal{L}^{MP}$	0.910 ± 0.008	0.888 ± 0.009
(A) w/o History	-	-	$\mathcal{L}^{MP}$	0.877 ± 0.010	0.847 ± 0.011
(B) w/o Attention	✓	Concat	$\mathcal{L}^{MP}$	0.893 ± 0.009	0.864 ± 0.010
(C) w/o MP-Loss	✓	Cross-Attn	Standard	0.900 ± 0.008	0.872 ± 0.010

Results are reported as mAUROC

large-scale fundus datasets, different patients frequently present with highly similar or identical pathological manifestations, making the standard one-hot negative assumption inappropriate. The Multi-Positive InfoNCE objective, by explicitly acknowledging semantic equivalence among similar image-text pairs, produces a more robust and well-calibrated joint embedding space.

Validity of the EHR signal and encoder design

To further examine the source of the gains attributed to EHR integration, we analyze the two additional settings in Table 4. First, randomly shuffling the EHR records across patients, Model E, collapses mAUROC to 0.864 and 0.835 on RFMiD and ODIR—below the no-history level Model A, 0.877/0.847—confirming that the model exploits the *patient-specific semantics* of the history rather than the mere presence of an extra text token acting as a regularizer. Second, sharing parameters between the two text encoders, Model D, reduces mAUROC by 1.1 and 1.8 percentage points relative to the full model, empirically justifying the dedicated dual-encoder design discussed in Sect. [Model architecture](#).

Discussion

Performance gains from EHR-augmented vision-language pre-training

The experimental results presented in Tables 1 and 2 demonstrate that **MedRecord-CLIP** delivers statistically significant gains over both vision-only foundation models and existing VLP frameworks on the multi-disease benchmarks (RFMiD and ODIR), and further leads on the IDRiD grading benchmark, while performing on par with the strongest baseline (RETFound) on APTOS 2019. The most notable gains are observed on RFMiD and ODIR, which feature dense multi-label annotations spanning a broad spectrum of retinal and systemic conditions. On RFMiD, our model improves mAUROC by 7.0 percentage points over FLAIR and 1.4 percentage points over CN-CLIP, demonstrating that EHR-augmented supervision provides richer semantic gradients than either standardized categorical text or general-domain image-text pairs. This pattern is consistent with the nature of the EHR signal: patient-level narratives are richest in systemic and cross-disease context, so their benefit is most pronounced precisely on the multi-disease benchmarks that depend on such context. Notably, this contextual supervision comes at no cost even on the DR-grading benchmarks, where severity is determined almost entirely by localized retinal lesions: MedRecord-CLIP still leads on IDRiD (DRS AUC 0.892 versus 0.867 for RETFound) and matches the strongest specialized baseline on APTOS 2019, showing that incorporating

EHR context strengthens multi-disease recognition without compromising performance on lesion-driven grading tasks. These findings validate the core premise of this work: that the semantic gap between rigid categorical labels and real-world clinical reasoning can be effectively bridged by incorporating patient-level EHR narratives as supervisory signals, most prominently in multi-disease settings with strong systemic correlates.

Clinical significance of individualized patient context

A central finding of our ablation study is that removing patient history text (Model A) produces the largest single-component performance drop across both benchmarks, with mAUROC decreasing by 3.3 and 4.1 percentage points on RFMiD and ODIR, respectively. This quantitative result carries direct clinical significance. In real-world ophthalmic practice, a senior clinician integrates fundus observations with a patient's systemic profile, including the duration of diabetes, blood pressure history, and presenting symptoms, to resolve visual ambiguities and detect co-occurring conditions. Vision-only models, by construction, lack access to this contextual layer, making them prone to under-diagnosing early-stage lesions whose visual signatures are subtle without systemic context. The pronounced performance gap on ODIR, which encompasses systemic conditions such as hypertension and diabetes, further substantiates this argument: diseases with strong systemic correlates benefit disproportionately from EHR integration, as the clinical history often encodes information that is not visually decodable from the fundus image alone. These observations align with the holistic diagnostic logic that motivates the construction of **MedRecordFundus** and suggest that EHR-aware VLP frameworks represent a clinically meaningful direction for next-generation fundus foundation models.

Architectural contribution of diagnosis-guided cross-attention

The ablation experiment comparing cross-attention fusion against naive concatenation (Model B) demonstrates that structured multimodal interaction is essential for effectively utilizing lengthy clinical narratives. Patient EHR records typically contain heterogeneous information of varying diagnostic relevance: a history of present illness entry may simultaneously describe systemic conditions, medication records, and lifestyle factors, only a subset of which bears directly on the current fundus findings. Simple concatenation of the diagnostic and history representations treats all components equally, introducing noise that weakens the discriminative capacity of

the fused feature. The diagnosis-guided cross-attention mechanism addresses this by positioning the diagnostic representation as a dynamic query that selectively attends to diagnostically salient segments of the history representation. The resulting 1.7 and 2.4 percentage point improvements in mAUC on RFMiD and ODIR confirm that adaptive, content-aware fusion is necessary to fully exploit the information embedded in individualized clinical records.

Robustness under semantic overlap via multi-positive contrastive learning

The comparison between Multi-Positive InfoNCE and standard CLIP contrastive loss (Model C) reveals a consistent improvement of 1.0 and 1.6 percentage points in mAUC on RFMiD and ODIR, respectively. While this margin is smaller than the gains from history integration or cross-attention fusion, its implications are methodologically important. Standard contrastive learning imposes a strict one-hot negative assumption, treating all unpaired samples within a batch as dissimilar. In large-scale medical datasets, this assumption is systematically violated: patients with the same diagnosis, comparable disease severity, and similar systemic profiles will produce image-text pairs that are semantically equivalent despite being nominally unpaired. Penalizing these pairs as negatives introduces conflicting gradient signals that erode the quality of the learned representations, a phenomenon particularly prevalent in fundus imaging where pathological manifestations of common conditions such as diabetic retinopathy are highly stereotyped. By reformulating the contrastive target as a multi-positive distribution, our objective aligns the model toward learning disease-level semantic consistency rather than instance-level memorization, producing a more robust and well-calibrated joint embedding space. This improvement in representation quality is reflected not only in the ablation results but also in the enhanced cross-dataset transferability observed on APTOS 2019 and IDRiD.

Limitations and future work

Despite the promising results, several limitations of the present study warrant acknowledgment. First, while **MedRecordFundus** contributes 21,290 fundus images and a professionally curated set of 1,043 individualized EHR narratives, its overall scale is relatively small, as assembling high-quality, de-identified records is labor-intensive; relative to large-scale non-medical VLP corpora, this leaves room to further enrich the diversity of clinical language patterns encountered during pre-training. Expanding the dataset to cover a broader range of systemic diseases, patient demographics, and imaging devices will be an important direction for future work.

Relatedly, the present scope of the EHR-bearing records leaves room for two complementary analyses that would further isolate the source of the EHR-driven gains. We do not pre-train a variant exclusively on **MedRecordFundus**, because a vision-language model trained from scratch on this in-house corpus alone would be constrained primarily by its data scale, so any resulting degradation could not be attributed cleanly to the absence of EHR rather than to the smaller pre-training corpus; we therefore isolate the EHR contribution through the shuffled-EHR control (Model E) instead. In addition, we do not report downstream performance broken down by clinical similarity to the EHR-bearing subset, since none of the four benchmarks provides EHR and this subset already spans the major sight-threatening conditions covered by the downstream tasks; a fine-grained, similarity-based breakdown is likewise left to future work. Second, the current framework operates on single-visit EHR snapshots and does not exploit longitudinal clinical trajectories, such as disease progression records across multiple consultations, which carry substantial prognostic value in conditions like diabetic retinopathy and glaucoma. Incorporating temporal EHR modeling represents a natural extension of the proposed framework. Third, although the diagnosis-guided cross-attention mechanism enhances interpretability relative to black-box fusion strategies, formal clinical validation through prospective studies and expert evaluation of attention attribution maps is necessary before deployment in real-world diagnostic workflows. Future work will also explore the integration of additional imaging modalities, such as Optical Coherence Tomography (OCT), to further enrich the multimodal representation and expand the clinical applicability of the framework.

Conclusion

In this work, we addressed a fundamental limitation shared by existing fundus vision-language pre-training frameworks: the systematic neglect of individualized patient context embedded in Electronic Health Records. We presented **MedRecordFundus**, a large-scale multimodal dataset comprising 21,290 CFP images and 1,043 expert-curated image-text pairs, each annotated across four structured clinical dimensions by senior ophthalmologists. To our knowledge, this is the first fundus dataset to incorporate individualized EHR narratives, spanning diagnosis, chief complaints, history of present illness, and past medical history, as supervisory signals for vision-language pre-training.

Leveraging this resource, we proposed **MedRecordCLIP**, a knowledge-enhanced foundation model that integrates a ViT-B/16 visual encoder with two RoBERTa-based text encoders through a novel **diagnosis-guided**

cross-attention mechanism. By positioning the diagnostic representation as a dynamic query over the patient history representation, the model adaptively retrieves diagnostically salient clinical evidence from lengthy EHR narratives, effectively suppressing irrelevant contextual noise. To further address the pervasive false-negative problem arising from semantic overlap among similar clinical descriptions, we introduced a symmetric **Multi-Positive InfoNCE objective** that extends the conventional one-hot contrastive target to a multi-positive formulation, fostering a more robust and well-calibrated joint embedding space.

Extensive experiments across four public benchmarks, namely RFMiD, ODIR, APTOS 2019, and IDRiD, demonstrate that **MedRecord-CLIP** achieves statistically significant improvements over state-of-the-art vision-only and vision-language baselines, including RETFound and FLAIR, on the multi-disease benchmarks (RFMiD and ODIR), and further leads on the IDRiD grading benchmark, while performing on par with the strongest baseline on APTOS 2019. Comprehensive ablation studies confirm that each proposed component, including clinical history integration, cross-attention fusion, and the multi-positive contrastive objective, contributes independently and complementarily to the overall performance gains. In particular, the substantial improvement observed upon removing patient history text underscores the clinical value of EHR context for resolving visual ambiguities in conditions with strong systemic correlates.

These results collectively demonstrate that bridging the semantic gap between fundus imagery and individualized clinical narratives is a promising and clinically grounded direction for advancing retinal disease diagnosis. We hope that **MedRecordFundus** and **MedRecord-CLIP** will serve as a foundation for future research on EHR-aware multimodal medical AI, with broader implications for population-scale screening, early-stage detection, and interpretable clinical decision support.

Acknowledgements

This work was supported in part by the Anhui Provincial Natural Science Foundation under Grant 2308085MF221, in part by the Training Project of Discipline (Major) Leaders in Anhui Universities under Grant DTR2025022, in part by the Anhui University Science and Engineering Teachers' Internship Practice Plan Project in Enterprises under Grant 2024jsqyg239, in part by the Open Research Project of Anhui Province Key Laboratory of Intelligent Building & Building Energy Saving of Anhui Jianzhu University under Grant IBES2025KF04ZD, and in part by the Exploratory Research Projects of Anhui University of Chinese Medicine under Grant AHUCM2024TS061 and AHUCM2024TS062.

Data availability

The public datasets (RFMiD, ODIR, APTOS 2019, IDRiD, FLAIR) analyzed during the current study are available in their respective public repositories. Owing to patient privacy and institutional data protection policies, a de-identified subset of MedRecordFundus and the pre-trained model checkpoints are available from the corresponding author on reasonable request.

Declarations

Conflict of interest

All authors declare that they have no conflict of interest.

Reproducibility statement

All reported results are averaged over five independent runs with different random seeds and presented as mean $\pm$ standard deviation, with the main comparisons evaluated via paired *t*-tests. The pre-training data preparation, the dataset splits, and the fine-tuning configuration are described in Sect. Experiments and results, while the corresponding source code and hyper-parameter configuration files will be released in our public repository at <https://github.com/calabashbro/MedRecord-CLIP-public> upon acceptance.

Author details

¹School of Computer and Information, Hefei University of Technology, Hefei 230009, Anhui, China. ²Anhui Provincial Key Laboratory of Intelligent Building and Building Energy Saving, Anhui Jianzhu University, Hefei 230022, Anhui, China. ³School of Medical Information Technology, Anhui University of Chinese Medicine, Hefei 230012, Anhui, China. ⁴The Second Affiliated Hospital, Anhui Medical University, Hefei 230601, Anhui, China.

Received: 2 April 2026 Accepted: 10 August 2026

Published online: 21 August 2026

References

1. Zhang J, Lin S, Cheng T, Xu Y, Lu L, He J, et al. Retfound-enhanced community-based fundus disease screening: real-world evidence and decision curve analysis. *NPJ Dig Med*. 2024;7(1):108.
2. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. *Nature*. 2023;622(7981):156–63.
3. Aghabeigi Aloooghareh MM, Sheikhey MM, Sahaifi A, Pirnejad H, Naemi A. Deep learning for comprehensive analysis of retinal fundus images: detection of systemic and ocular conditions. *Bioengineering*. 2025;12(8):840.
4. Qiu J, Wu J, Wei H, Shi P, Zhang M, Sun Y, Li L, Liu H, Liu H, Hou S, et al. Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. *arXiv preprint arXiv:2310.04992*. 2023.
5. Wang Z, Wu Z, Agarwal D, Sun J. Medclip: Contrastive learning from unpaired medical images and text. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. 2022; 3876.
6. Silva-Rodriguez J, Chakor H, Kobbi R, Dolz J, Ayed IB. A foundation language-image model of the retina (flair): encoding expert knowledge in text supervision. *Med Image Anal*. 2025;99:103357.
7. Wu R, Zhang C, Zhang J, Zhou Y, Zhou T, Fu H. Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2024; 722–732.
8. Messica S, Cohen S, Hadad A, Gordon M, Katz O, Presil D, et al. Temporal integrative machine learning for early detection of diabetic retinopathy using fundus imaging and electronic health records. *IEEE J Biomed Health Inform*. 2025;30(6):4622–33.
9. Shi D, Zhang W, Yang J, Huang S, Chen X, Xu P, et al. A multimodal visual-language foundation model for computational ophthalmology. *NPJ Dig Med*. 2025;8(1):381.
10. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. 2021;8748–63.
11. Du J, Guo J, Zhang W, Yang S, Liu H, Li H, et al. Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer; 2024;709–19.

12. Pachade S, Porwal P, Thulkar D, Kokare M, Deshmukh G, Sahasrabudhe V, et al. Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research. *Data*. 2021;6(2):14.
13. University, P. Ocular Disease Intelligent Recognition (ODIR-5K). <https://odir2019.grand-challenge.org/> 2019.
14. Society AP, T-O. APTOS 2019 Blindness Detection. Kaggle; 2019.
15. Porwal P, et al. Idrid: diabetic retinopathy-segmentation and grading challenge. *Med Image Anal*. 2020;59:101561.
16. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(1):87–110.
17. Bannur S, Hyland S, Liu Q, Perez-Garcia F, Ilse M, Castro DC, et al. Learning to exploit temporal structure for biomedical vision-language processing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023;15016–27.
18. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. *Adv Neural Inf Process Syst*. 2023;36:28541–64.
19. Yang A, Pan J, Lin J, Men R, Zhang Y, Zhou J, Zhou C. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335* 2022.
20. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. 2019;6105–14 (**PMLR**).
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*. 1982;143(1):29–36.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.