Towards Zero-Shot Diabetic Retinopathy Grading: Learning Generalized Knowledge via Prompt-Driven Matching and Emulating

Huan Wang[1], Haoran Li[1], Yuxin Lin[2], Huaming Chen[3], Jun Yan[1], Lijuan Wang[4], Jiahua Shi[5], Qihao Xu[2], Yongting Hu[2], Yong Xu[2*], Jun Shen[1*]

[1]School of Computing and Information Technology, University of Wollongong, Australia
[2]Harbin Institute of Technology, ShenZhen, China
[3]School of Electrical and Computer Engineering, The University of Sydney, Australia
[4]School of Cyber Engineering, Xidian University, Xi'an, China
[5]The University of Queensland, Australia
hw226@uowmail.edu.au, laterfall@hit.edu.cn, jshen@uow.edu.au

Abstract

As one of the primary causes of visual impairment, Diabetic Retinopathy (DR) requires accurate and robust grading to facilitate timely diagnosis and intervention. Different from conventional DR grading methods that utilize single-view images, recent clinical studies have revealed that multi-view fundus images can significantly enhance DR grading performance by expanding the field of view (FOV). However, there is a long-tailed distribution problem in fundus image analysis, *i.e.*, a high prevalence of mild DR grades and a low prevalence of rare ones (*e.g.*, cases of high severity), which presents a significant challenge to developing a unified model capable of detecting rare or unseen DR grades not encountered during training. In this paper, we propose **ProME-DR**, a **Pro**mpt-driven zero-shot **DR** grading framework, which leverages prompt **M**atching and **E**mulating to recognize the unseen DR categories and views beyond the training set. ProME-DR disentangles the training process into two stages to learn generalized knowledge for novel DR disease grading. Initially, ProME-DR leverages two sets of prompt units to capture semantic and inter-view consistency knowledge via a split-and-mask manner, gathering instance-level DR visual clues. Subsequently, it constructs a concept-aware emulator to generate context prompt units, linking extensible knowledge learned from the previously seen DR attributes for zero-shot DR grading. Extensive experiments conducted on eight datasets and various scenarios confirm the superiority of ProME-DR.

Code — https://github.com/hwang52/ProMEDR

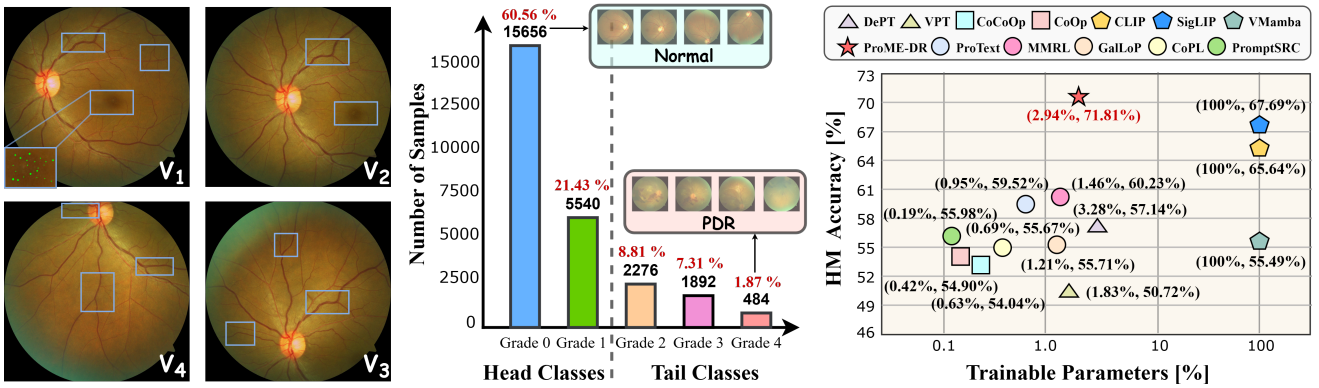
Introduction

According to the International Diabetes Federation (IDF) (Sun et al. 2022), 537 million individuals globally were affected by diabetes, and projections indicate a surge in diabetes to 783 million by 2045. Diabetic Retinopathy (DR) (Teo et al. 2021) is supposed to be a leading cause of visual impairment and blindness. As an adverse complication of diabetes, early detection of DR severity grades is crucial to decrease its prevalence. Following (Heng et al. 2013), the severity levels of DR can be divided into five stages (grade 0-4) from light to serious: *normal, mild, moderate, severe, and Proliferative Diabetic Retinopathy (PDR)*. To prevent and screen for diabetic retinopathy, developing accurate and robust DR grading models is an invaluable asset in enabling earlier treatments (Atwany, Sahyoun, and Yaqub 2022; Dai et al. 2021).

Recently, with the successful development of deep learning techniques (Suzuki 2017; Kumar et al. 2024), a growing number of studies have explored DR grading tasks with fundus images (Dai et al. 2024). However, most existing DR grading methods still fall short of achieving high-precision diagnosis (Zhang et al. 2022). The main reason is that most of the previous works are trained on the single-view databases (*e.g.*, DeepDR (Liu et al. 2022a) and DDR (Li et al. 2019)) with a field of view (FOV) of only $45°$-$50°$. Actually, the human eye's FOV is about $200°$-$220°$, which means that single-view fundus images may lead to loss of most of the lesion information on retinal areas (Hu et al. 2019). As shown in Fig. 1 (a), taking the latest multi-view DR image dataset MFIDDR (Luo et al. 2023) as an example, a single-view image only obtains limited pathological information, *e.g.*, the hard exudates are only visible in V_1 (top-left) and V_2 (top-right) views. Previous works (Luo et al. 2023, 2025; Lin et al. 2025; Luo et al. 2021) have revealed that multi-view methods have promising accuracy for DR grading by incorporating complementary information from multiple DR fundus views.

On the other hand, regardless of single- or multi-view DR grading, a consistent challenge stems from a long-tailed distribution problem, characterized by heavily imbalanced datasets. For example, as shown in Fig. 1 (b), the normal and PDR cases in MFIDDR (Luo et al. 2023) account for 60.56% and 1.87% of the training set. In practice, collecting the standard annotation for every DR grade from clinical experts can be highly expensive and may raise privacy concerns (Padhy et al. 2019; Ursin et al. 2021), and the cases of high severity are typically underrepresented. In such a scenario, a zero-shot DR grading approach is highly desired, where the model can automatically recognize unseen DR grades without prior exposure to annotated DR samples during training. Hence, we aim to explore the potential of zero-shot learning in developing a generalized DR detector.

*Corresponding Authors: Yong Xu, Jun Shen.



(a) Multi-view fundus images (b) Proportion of each grade in MFIDDR (c) Seen-to-Unseen comparison results

Figure 1: **(a)** An example of multi-view fundus images from MFIDDR (Luo et al. 2023). **(b)** Long-tailed distribution of DR classes in fundus image datasets (taking MFIDDR as an example). **(c)** Comparison of our proposed ProME-DR with competing methods, showing the average Harmonic Mean (HM) accuracy over eight DR datasets (see details in Table 1).

The mainstream zero-shot classification studies (Lai et al. 2024; Martin et al. 2024) pay attention to fine-tuning foundation models (*e.g.*, CLIP (Radford et al. 2021)), under the guidance of expert knowledge to enhance the accuracy of downstream tasks. Such compelling representational capability is conducive to the generalization of the models for unseen images. However, conventional fine-tuning approaches inevitably lead to forgetting the prior knowledge acquired during large-scale pre-training (Bafghi et al. 2024; Bai et al. 2025), and also incur significant computational costs (Xia et al. 2024). To alleviate this problem, we explore the use of prompt tuning (Liu et al. 2023; Xing et al. 2024), which is a more flexible and lightweight paradigm than fine-tuning. However, when we directly apply prompt tuning methods for zero-shot DR grading, such as visual prompt tuning (VPT) (Jia et al. 2022), we observe these methods indeed offer modest computational savings but exhibit no clear advantage in the DR performance, and in some cases, they even underperform compared to the fine-tuned CLIP, as shown in Fig. 1 (c). This may be attributed that 1) they fail to capture the unique properties of DR fundus images, resulting in suboptimal generalization, and 2) the visual clues of the semantics for DR fundus images are usually more complex and ambiguous than natural images.

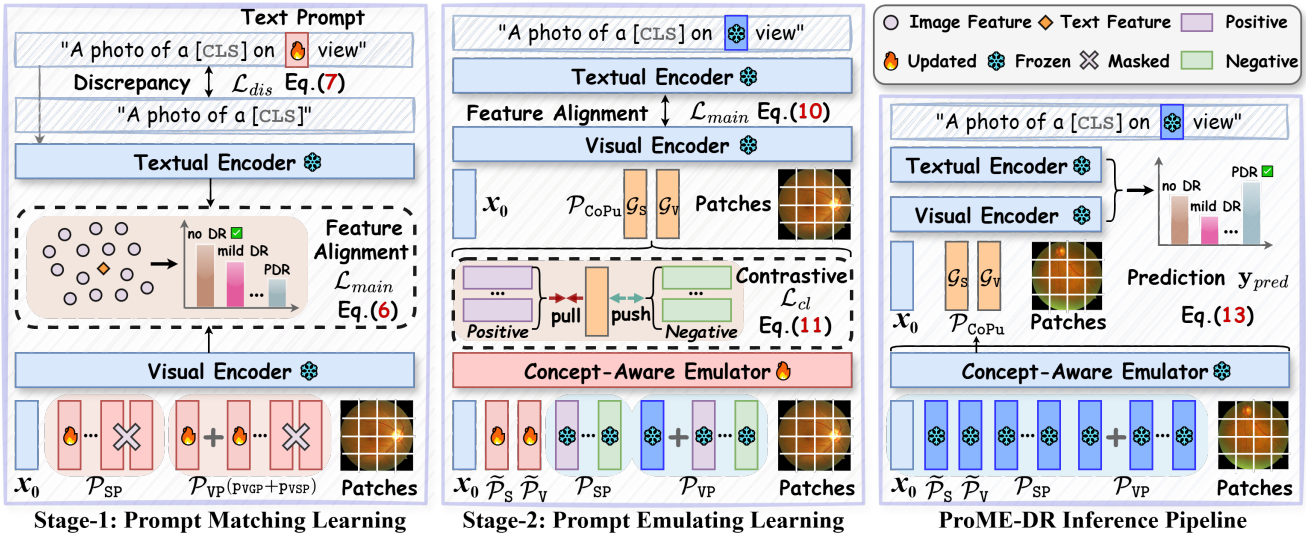
In this paper, we propose **ProME-DR**, a **Pro**mpt-driven framework for zero-shot **DR** grading with multi-view fundus images, which learns generalized knowledge from CLIP via prompt **M**atching and **E**mulating to recognize the unseen DR categories and views. Specifically, ProME-DR decouples the learning process into two stages, including Prompt Matching Learning (**PML** in Stage-**1**) and Prompt Emulating Learning (**PEL** in Stage-**2**), allowing the model to first comprehensively understand the visual features of DR fundus images from various views, and then focus on simulating properties of the unseen sample for zero-shot DR grading. In PML, we devise two sets of distinct prompt units: semantic and view prompts, which are employed to extract class-relevant and inter-view consistency knowledge from the training set in a split-and-mask strategy. In PEL, we then construct a Concept-Aware Emulator (CAE) under contrastive constraint signals to dynamically produce Context Prompt units (CoPu) based on the learned prompt units. With the context prompt units, facing agnostic scenarios, we can leverage the extensive knowledge from CLIP to obtain generalized representations for substantially enhanced zero-shot DR grading, as shown in Fig. 1 (c). Our main contributions are summarized as follows:

- We propose ProME-DR, which, to our best knowledge, is the first framework established for zero-shot DR grading. It enables the model to obtain fine-grained visual features that can effectively generalize to unseen DR samples.
- We introduce two novel components in ProME-DR, including PML and PEL, to implement a two-stage prompt-driven approach for zero-shot DR grading with the multi-view fundus images. PML learns the generalized knowledge of DR via a prompt-matching method, while PEL uses a prompt-emulating method to deal with scenarios of unseen DR categories and views.
- We further establish a comprehensive benchmark for the zero-shot DR grading task on eight DR fundus image datasets. Our ProME-DR is evaluated in multiple DR evaluation scenarios in this benchmark and consistently outperforms state-of-the-art counterparts.

Related Work

Diabetic Retinopathy Grading

With the development of deep learning algorithms (He et al. 2016; Han et al. 2022), there are many methods widely used for the DR grading task, which can be divided into two directions in terms of architecture: CNN-based and Transformer-based. (Kalyani et al. 2023) implemented an automatic diagnosis of DR using fundus image detection with transfer learning. (Luo et al. 2024) proposed a novel DR grading framework to jointly learn fundus images from multiple fields of view. (Dai et al. 2023) uses a twin transformer



Stage-1: Prompt Matching Learning **Stage-2: Prompt Emulating Learning** **ProME-DR Inference Pipeline**

Figure 2: **Overview of our proposed ProME-DR**. **Left**. In Stage-1: Prompt Matching Learning (PML), we propose semantic prompts $\mathcal{P}_{\text{SP}}$ and view prompts $\mathcal{P}_{\text{VP}}$, to capture fine-grained visual features from DR categories and views via a split-and-mask strategy. **Middle**. In Stage-2: Prompt Emulating Learning (PEL), we introduce a concept-aware emulator to dynamically generate context prompt units $\mathcal{P}_{\text{CoPu}}$. After training, the generated $\mathcal{P}_{\text{CoPu}}$ can be aware of visual clues related to the DR attributes in an unknown scenario. **Right**. The inference pipeline (Eq. (13)) of our proposed ProME-DR.

to capture both global and local features of fundus images to enhance performance. Recently, based on the multi-view fundus image dataset (Luo et al. 2023), multi-view learning (Yu et al. 2025) offers fresh opportunities for DR grading. (Luo et al. 2023) proposed a multi-view model for DR detection by fusing local features and long-range global features. Then, (Lin et al. 2025) introduced a multi-view framework based on the learnable retinal vessel reinforcement block. (Luo et al. 2025) further devised a dynamic selection-driven multi-view DR grading method to suit clinical scenarios better. Nevertheless, these methods suffer from the long-tailed distribution issue, and they tend to overfit to frequent categories and views in the training set. In addition, facing unseen categories and views beyond the training set, they are almost unrecognizable. In contrast, our ProME-DR takes the first step toward zero-shot DR grading by enabling generalization with unseen DR samples.

Prompt Tuning with Pre-trained Models

Prompt Tuning (Min et al. 2023; Liu et al. 2023), as a popular and efficient tuning paradigm, originally appeared in natural language processing for adapting pre-trained language models to various downstream datasets and tasks (Min et al. 2023; Wu et al. 2024). CLIP (Radford et al. 2021) efficiently learns the visual concepts from the large-scale image-text pairs in a contrastive learning way. Considering the efficiency of parameter size and convergence rate, prompt learning is rapidly expanding to adapt pre-trained models (Khattak et al. 2023; Li et al. 2024). CoPL (Goswami et al. 2024) or CoCoOp (Zhou et al. 2022a) fine-tuned the CLIP model by optimizing a continuous set of prompt vectors. ProS (Fang et al. 2024) introduces prompting-to-simulate to apply prompt tuning for UCDR, employing a two-step process to simulate dynamic prompts that can impact models to produce generalized features. However, these methods are static after training, as their learned prompts are fixed and potentially overfit to the visual features of the frequent seen DR grades. Differently, our ProME-DR adopts a two-stage framework of matching first and then emulating to handle the zero-shot DR grading task.

Methodology

We make the first attempt to apply CLIP with prompt tuning for the zero-shot DR grading task, and propose a novel two-stage prompt-driven framework, ProME-DR, as illustrated in Fig. 2. First, we briefly outline the preliminaries about CLIP and the prompt tuning paradigm. Then, we provide a detailed description of ProME-DR, which comprises two novel components: Prompt Matching Learning (PML) and Prompt Emulating Learning (PEL). Finally, we summarize the overall training and inference pipelines of ProME-DR.

Preliminaries

Contrastive Language-Image Pre-Training CLIP (Radford et al. 2021) initially indicates that, equipped with large-scale image-text pairs pre-training, a contrastive learning-based foundation model can achieve comparable or even surpass performance with a fully supervised method. Concretely, CLIP aligns the texts and images through an image encoder $f_i(\cdot)$ and a textual encoder $f_t(\cdot)$. CLIP performs the zero-shot classification based on the similarity between the visual features $f_i(x)$ (sample x) and text features. The text features of captions $\{\mathtt{t}_{y_1}, \ldots, \mathtt{t}_{y_{|\mathcal{C}|}}\}$ from different categories (y as label, $|\mathcal{C}|$ is the number of categories) as $\{f_t(\mathtt{t}_{y_j})\}_{j=1}^{|\mathcal{C}|}$, which are produced by adding the class

names into a text template. For example, $\mathtt{t}_{y_j}$ is the caption of class y_j as 'a photo of a $[\mathtt{cls}]^j$' where $[\mathtt{cls}]^j$ is the class name of y_j (*e.g.*, $[\mathtt{cls}]^j \rightarrow$ 'cat'). Given an image x and different categories' captions, CLIP outputs a prediction by:

$$\mathbf{y}_{clip} = \arg\max_j (f_i(x) \otimes f_t(\mathtt{t}_{y_j})), \tag{1}$$

where $\otimes$ is the cosine similarity operator and j is the label index. Based on Eq. (1), CLIP can classify images according to the candidate class names in a zero-shot manner. Owing to the large-scale contrastive pre-training of images and texts (You et al. 2022), CLIP exhibits remarkable zero-shot performance on diverse downstream tasks (Fang et al. 2022). Therefore, we focus on how to effectively leverage the generalized knowledge embedded in a pre-trained CLIP for the zero-shot DR grading task.

Prompt Tuning Paradigm Prompt tuning adapts frozen pre-trained models to different downstream tasks by introducing tunable parameters to the input space. Specifically, CoOp (Zhou et al. 2022b) exploits soft-prompting ideas to train dynamic learnable prompt vectors $\mathcal{P}_{tex}^{M_{tex}} = \{\mathtt{p}_{tex}^l \in \mathbb{R}^\ell\}_{l=1}^{M_{tex}}$ and preserve the semantic relationship between textual concepts and labels, where M_{tex} and $\mathbb{R}^\ell$ mean text prompts length and dimensions. The text template of CoOp can be denoted as the '$\mathcal{P}_{tex}^{M_{tex}}$ a photo of a $[\mathtt{cls}]$', and it is then fed into the CLIP's text encoder. CoCoOp (Zhou et al. 2022a) improves the performance of CoOp by generating prompts conditioned on each input image, which is formulated as $\mathcal{P}_{tex}^{M_{tex}} = \{\mathtt{p}_{tex}^l + \Phi(f_i(x))\}_{l=1}^{M_{tex}}$, where Φ refers to the meta-net and x is the input image. Furthermore, VPT (Jia et al. 2022) constructs visual prompts for ViT (Dosovitskiy et al. 2020) as $\mathcal{P}_{vis}^{M_{vis}} = \{\mathtt{p}_{vis}^l \in \mathbb{R}^d\}_{l=1}^{M_{vis}}$, where M_{vis} means the visual prompts length and $\mathbb{R}^d$ denotes the visual prompts dimensions. Concretely, VPT-Shallow (Jia et al. 2022) inserts the visual prompts into the first ViT's transformer layer as:

$$\begin{aligned}
[\mathbf{x}_1, \mathbf{Z}_1, \mathbf{E}_1] &= L_1([\mathbf{x}_0, \mathcal{P}_{vis}^{M_{vis}}, \mathbf{E}_0]), \\
[\mathbf{x}_n, \mathbf{Z}_n, \mathbf{E}_n] &= L_n([\mathbf{x}_{n-1}, \mathbf{Z}_{n-1}, \mathbf{E}_{n-1}]),\ n = 2, 3, ..., N, \\
\mathbf{y}_{output} &= \mathtt{Head}(\mathbf{x}_N),
\end{aligned} \tag{2}$$

where $\mathbf{x}_n \in \mathbb{R}^d$ is the embedding of the $[\mathtt{cls}]$ token at L_{n+1} layer's input space, $\mathbf{E}_0$ means the embedding of the input image patches, and $\mathbf{Z}_n$ represents the embedding of visual prompts $\mathcal{P}_{vis}^{M_{vis}}$ calculated by L_n transformer layer.

Stage-1: Prompt Matching Learning (PML)

In Prompt Matching Learning (PML), we first introduce two sets of distinct prompt units: semantic and view prompts, to capture class-relevant and inter-view consistency knowledge during PML, as illustrated in Fig. 2-**Left**. Concretely, the Semantic Prompts $\mathcal{P}_{\mathtt{SP}}$ consist of $|\mathcal{C}|$ learnable vectors to focus on the category-specific representations:

$$\mathcal{P}_{\mathtt{SP}} = \{\mathtt{p}_{\mathtt{SP}}^l \in \mathbb{R}^d\}_{l=1}^{|\mathcal{C}|}. \tag{3}$$

Then, for the View Prompts $\mathcal{P}_{\mathtt{VP}}$, to consider inherent correlations from multiple views, we disentangle $\mathcal{P}_{\mathtt{VP}}$ and assign it as view-general prompts and view-specific prompts:

$$\mathcal{P}_{\mathtt{VP}} = \{\mathtt{p}_{\mathtt{VP}}^l = (\mathtt{p}_{\mathtt{VGP}} + \mathtt{p}_{\mathtt{VSP}}^l)\}_{l=1}^{|\mathcal{V}|}, \tag{4}$$

where dimensions of view-general prompts $\mathtt{p}_{\mathtt{VGP}}$ and view-specific prompts $\mathtt{p}_{\mathtt{VSP}}^l$ are $\mathbb{R}^d$, and $|\mathcal{V}|$ is the number of views. Based on Eq. (4), $\mathtt{p}_{\mathtt{VGP}}$ captures shared information to extract inter-view invariant knowledge across different views, while $\mathtt{p}_{\mathtt{VSP}}$ gathers specific intra-view discriminative knowledge from the corresponding DR view.

To ensure that each prompt focuses on fine-grained visual clues related to a corresponding category or view, we only explicitly optimize single $\mathtt{p}_{\mathtt{SP}}$ and $\mathtt{p}_{\mathtt{VP}}$ in PML. We construct binary mask matrix $\mathcal{M}_{\mathtt{SP}} = \{\mathtt{m}_{\mathtt{SP}}^l \in \{0, 1\}\}_{l=1}^{|\mathcal{C}|}$ and $\mathcal{M}_{\mathtt{VP}} = \{\mathtt{m}_{\mathtt{VP}}^l \in \{0, 1\}\}_{l=1}^{|\mathcal{V}|}$ to mask irrelevant semantic and view prompts, where $\mathtt{m}_{\mathtt{SP}}^l$ and $\mathtt{m}_{\mathtt{VP}}^l$ denotes a binary variable. For example, the input sample with class c and view v, the $\mathtt{m}_{\mathtt{SP}}^c$ and $\mathtt{m}_{\mathtt{VP}}^v$ are 1 and others are 0. Subsequently, the masked semantic and view prompts are inserted into the input space of the CLIP's image encoder with other tokens:

$$\mathcal{I}_{\mathtt{PML}} = [\mathbf{x}_0, \mathcal{M}_{\mathtt{SP}} \circ \mathcal{P}_{\mathtt{SP}}, \mathcal{M}_{\mathtt{VP}} \circ \mathcal{P}_{\mathtt{VP}}, \mathbf{E}_0], \tag{5}$$

where $\circ$ denotes the element-wise product operator. The embedding of the $[\mathtt{cls}]$ token is obtained based on Eq. (2). Then, $\mathcal{I}_{\mathtt{PML}}$ is fed into the first ViT's transformer layer.

The text features serve as anchors to attract neighboring images, benefiting from CLIP's robust image-text matching capabilities. To facilitate more flexible learning, we introduce a new textual template for DR grading with multi-view fundus images, as $\mathcal{T}_{y_j}$ = 'a photo of a $[\mathtt{cls}]^j$ on $\mathcal{T}_{\mathtt{DR}}$ view', where $[\mathtt{cls}]^j$ is the name of DR grade y_j (*e.g.*, PDR), $\mathcal{T}_{\mathtt{DR}} \in \mathbb{R}^{K \times \ell}$ denotes a set of learnable text prompts with prompts length K. Formally, given the fused input $\mathcal{I}_{\mathtt{PML}}$ and DR grades' captions $\{\mathcal{T}_{y_j}\}_{j=1}^{|\mathcal{C}|}$ generated by the above text template, the training objective can be formulated as:

$$\mathcal{L}_{main} = \sum_{j=1}^{|\mathcal{C}|} -\mathbf{1}_{y_j} \log(f_i(\mathcal{I}_{\mathtt{PML}}) \otimes f_t(\mathcal{T}_{y_j})), \tag{6}$$

where $\otimes$ is the cosine similarity operator, $f_i(\cdot)$ and $f_t(\cdot)$ are the CLIP's image and text encoders, and j is the label index.

Considering that mask training in PML might lead to overfitting and forgetting of essential linguistic knowledge about DR. Inspired by (Khattak et al. 2023), we further narrow down the discrepancy between learnable text prompts $\mathcal{T}_{y_j}$ and original CLIP's text prompts $\mathtt{t}_{y_j}$, it can be formulated as the following objective:

$$\mathcal{L}_{dis} = \sum_{\mathrm{dim}=1}^{\ell} \|f_t(\mathcal{T}_{y_j}) - f_t(\mathtt{t}_{y_j})\|_2^2, \quad f_t(\mathcal{T}_{y_j}) \in \mathbb{R}^\ell, \tag{7}$$

where $\mathtt{t}_{y_j}$ is the caption of class y_j as 'a photo of a $[\mathtt{cls}]^j$'. Based on Eq. (7), $\mathcal{L}_{dis}$ guide the learnable $\mathcal{T}_{y_j}$ to gain complementary knowledge from pre-trained CLIP text features. Overall, the training objective of PML as $\mathcal{L}_{stage1}$:

$$\mathcal{L}_{stage1} = \mathcal{L}_{main} + \lambda_1 \cdot \mathcal{L}_{dis}, \tag{8}$$

where λ_1 represents the loss balancing factor. In the Stage-1 PML, we perform image-text alignment training via Eq. (8) based on the masked semantic and view prompts, gathering fine-grained visual features from DR samples.

Stage-2: Prompt Emulating Learning (PEL)

In Prompt Emulating Learning (PEL), we aim to refine a set of Context Prompt units (CoPu, $\mathcal{P}_{\texttt{CoPu}} = [\mathcal{G}_{\texttt{S}} \in \mathbb{R}^d, \mathcal{G}_{\texttt{V}} \in \mathbb{R}^d]$) for DR grading and ensure that $\mathcal{G}_{\texttt{S}}$ and $\mathcal{G}_{\texttt{V}}$ can generalize to the categories and view attributes of unseen DR samples. The key insight is to endow $\mathcal{P}_{\texttt{CoPu}}$ with the ability to capture extensive knowledge from CLIP about unknown DR samples, by linking the $\mathcal{P}_{\texttt{SP}}$ and $\mathcal{P}_{\texttt{VP}}$ that are trained in Stage-1 PML. Specifically, we design a Concept-Aware Emulator (CAE) to dynamically generate $\mathcal{P}_{\texttt{CoPu}}$ based on a fused input $\mathcal{I}_{\texttt{PEL}}$. Given the input sample x, the fused input $\mathcal{I}_{\texttt{PEL}}$ can be formulated as:

$$\mathcal{I}_{\texttt{PEL}} = [\mathbf{x}_0, \widetilde{\mathcal{P}}_{\texttt{S}}, \widetilde{\mathcal{P}}_{\texttt{V}}, \mathcal{P}_{\texttt{SP}}, \mathcal{P}_{\texttt{VP}}, \mathbf{E}_0], \tag{9}$$

where $\widetilde{\mathcal{P}}_{\texttt{S}} \in \mathbb{R}^d$ and $\widetilde{\mathcal{P}}_{\texttt{V}} \in \mathbb{R}^d$ are used to simulate semantic and view properties of the environment, as CAE's inputs. Let h_θ denotes the CAE parameterized by θ, we use same objective in Eq. (6) to optimize $\widetilde{\mathcal{P}}_{\texttt{S}}$, $\widetilde{\mathcal{P}}_{\texttt{V}}$, and θ:

$$\begin{aligned} \mathcal{L}_{main} &= \sum_{j=1}^{|\mathcal{C}|} -\mathbf{1}_{y_j} \log(f_i(\mathcal{I}_{context}) \otimes f_t(\mathcal{T}_{y_j})) \\ \text{s.t.} \quad & \mathcal{I}_{context} = [\mathbf{x}_0, h_\theta(\mathcal{I}_{\texttt{PEL}}), \mathbf{E}_0], \end{aligned} \tag{10}$$

where $\mathcal{P}_{\texttt{CoPu}} = [\mathcal{G}_{\texttt{S}}, \mathcal{G}_{\texttt{V}}] = h_\theta(\mathcal{I}_{\texttt{PEL}})$, and $\mathcal{T}_{y_j}$ remains frozen and is trained in Stage-1 only. To allow generated $\mathcal{G}_{\texttt{S}}$ and $\mathcal{G}_{\texttt{V}}$ to concentrate on more relevant visual information, inspired by contrastive learning (Chen et al. 2020; He et al. 2020; Grill et al. 2020), we regard $\texttt{p}^c_{\texttt{SP}}$ and $\texttt{p}^v_{\texttt{VP}}$ with the corresponding class c and view v as the positive pair, and other prompt units as the negative pairs. We design an objective term $\mathcal{L}_{cl}$ to maximize the representation similarity between the query and positive pairs, and minimize the similarity between negative pairs. It is natural to derive the following optimization objective term:

$$\begin{aligned} \mathcal{L}_{cl} = -\frac{1}{2}(& \log \frac{\exp(s(\mathcal{G}_{\texttt{S}}, \texttt{p}^c_{\texttt{SP}})/\tau)}{\sum_{\texttt{p}^j_{\texttt{SP}} \in \mathcal{N}_{\texttt{SP}}} \exp(s(\mathcal{G}_{\texttt{S}}, \texttt{p}^j_{\texttt{SP}})/\tau)} \\ & + \log \frac{\exp(s(\mathcal{G}_{\texttt{V}}, \texttt{p}^v_{\texttt{VP}})/\tau)}{\sum_{\texttt{p}^j_{\texttt{VP}} \in \mathcal{N}_{\texttt{VP}}} \exp(s(\mathcal{G}_{\texttt{V}}, \texttt{p}^j_{\texttt{VP}})/\tau)}) \\ \text{s.t.} \quad & s(a,b) = \frac{a \cdot b}{\|a\|_2 \times \|b\|_2}, \end{aligned} \tag{11}$$

where $\mathcal{N}_{\texttt{SP}} = \{\mathcal{P}_{\texttt{SP}} \backslash \texttt{p}^c_{\texttt{SP}}\}$, $\mathcal{N}_{\texttt{VP}} = \{\mathcal{P}_{\texttt{VP}} \backslash \texttt{p}^v_{\texttt{VP}}\}$, and τ controls the concentration strength of the representations (Chen et al. 2020). Therefore, we expect to enlarge the similarity with attribute coincident prompt units than others, encouraging associations with previously seen DR visual clues. The whole training objective of PEL can then be defined as $\mathcal{L}_{stage2}$:

$$\mathcal{L}_{stage2} = \mathcal{L}_{main} + \lambda_2 \cdot \mathcal{L}_{cl}, \tag{12}$$

where λ_2 denotes the loss balancing factor for $\mathcal{L}_{cl}$. Notably, in Stage-2, only the $\widetilde{\mathcal{P}}_{\texttt{S}}$, $\widetilde{\mathcal{P}}_{\texttt{V}}$, and θ are trained, while other prompts remain frozen, as shown in Fig. 2-**Middle**. After the PEL training, based on context prompt units $\mathcal{P}_{\texttt{CoPu}}$, the refined $\mathcal{G}_{\texttt{S}}$ and $\mathcal{G}_{\texttt{V}}$ can be aware of fine-grained visual features related to the DR attributes in an unknown scenario, thereby achieving zero-shot DR grading.

Pipelines of ProME-DR

In summary, we perform a two-stage training pipeline in our ProME-DR: Prompt Matching Learning (PML) and Prompt Emulating Learning (PEL). In Stage-1, we optimize the $\mathcal{P}_{\texttt{SP}}$ and $\mathcal{P}_{\texttt{VP}}$ in PML to capture class-relevant and inter-view consistency knowledge about DR fundus images, while tuning text prompts $\mathcal{T}_{\texttt{DR}}$ designed for DR grading. In Stage-2, we train the CAE h_θ to dynamically generate $\mathcal{P}_{\texttt{CoPu}} = [\mathcal{G}_{\texttt{S}}, \mathcal{G}_{\texttt{V}}]$ based on CAE's input $\widetilde{\mathcal{P}}_{\texttt{S}}$ and $\widetilde{\mathcal{P}}_{\texttt{V}}$ in PEL. The generated $\mathcal{G}_{\texttt{S}}$ and $\mathcal{G}_{\texttt{V}}$ can effectively impact CLIP to discover relevant visual clues, thereby generalizing to unseen DR samples. We impose $\mathcal{L}_{stage1}$ in Eq. (8) and $\mathcal{L}_{stage2}$ in Eq. (12) to supervise the training in PML and PEL. As shown in Fig. 2-**Right**, the inference of ProME-DR is as follows:

$$\begin{aligned} & \mathcal{G}_{\texttt{S}}, \mathcal{G}_{\texttt{V}} = h_\theta([\mathbf{x}_0, \widetilde{\mathcal{P}}_{\texttt{S}}, \widetilde{\mathcal{P}}_{\texttt{V}}, \mathcal{P}_{\texttt{SP}}, \mathcal{P}_{\texttt{VP}}, \mathbf{E}_0]), \\ & \mathcal{I}_{context} = [\mathbf{x}_0, \mathcal{G}_{\texttt{S}}, \mathcal{G}_{\texttt{V}}, \mathbf{E}_0], \\ & \mathbf{y}_{pred} = \arg\max_j (f_i(\mathcal{I}_{context}) \otimes f_t(\mathcal{T}_{y_j})), \end{aligned} \tag{13}$$

where $j \in \{1, \ldots, |\mathcal{C}|\}$ and $\otimes$ is a cosine similarity operator. The full algorithm is provided in Appendix A.

Experiments

Experimental Setup

Datasets. We evaluate the effectiveness of our ProME-DR on 8 standard DR datasets, including multi-view MFIDDR (Luo et al. 2023) and the single-view GDR-Bench (Che et al. 2023) datasets: APTOS (Karthik, Maggie, and Dane 2019), DeepDR (Liu et al. 2022b), IDRiD (Porwal et al. 2018), MESSIDOR (Abràmoff et al. 2013), RLDR (Wei et al. 2021), DDR (Li et al. 2019), and EyePACS (Gulshan et al. 2016). The MFIDDR comprises $8,613$ samples, and each associated with 4 fundus images ($V_1 \sim V_4$): V_1 focuses on the macula, V_2 centers on the optic disc, V_3 and V_4 are tangent to upper and lower horizontal lines of the optic disc.

Scenarios. We validate ProME-DR in three scenarios: **1)** Seen-to-Unseen: following (Zhou et al. 2022a), we randomly sample for each dataset a few-shot training set while using the original test set for testing. We evaluate the highest shot number studied in (Zhou et al. 2022b) (*i.e.*, 16 shots). Considering that the class 0 (normal) is the most common grade in DR, we set class 0 as the seen class, and other classes (DR grade $1 \sim 4$) as the unseen classes. Similarly, for the multi-view dataset MFIDDR (Luo et al. 2023), we set view V_1 as the seen view, and other views (view $V_2 \sim V_4$) as the unseen views. **2)** Cross-Dataset Generalization: we set the MFIDDR dataset (Luo et al. 2023) as the source dataset and other 7 DR datasets as target datasets for direct recognition. **3)** Supervised Evaluation: we compare with state-of-the-art DR grading methods on the MFIDDR (Luo et al. 2023) dataset with supervised learning, including the single-view and multi-view DR grading methods.

Compared Methods. We compare ProME-DR with four types of baselines: 1) zero-shot prompt-based: CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), VPT (Jia et al. 2022), PromptSRC (Khattak et al. 2023), GalLoP (Lafon

Compared Methods	Protocols	MFIDDR	APTOS	DeepDR	IDRiD	MESSIDOR	RLDR	DDR	EyePACS	**Average**
CoOp [IJCV22]	Seen	88.50	92.30	72.44	74.12	64.92	73.62	87.39	82.78	79.51
	Unseen	37.14	48.93	30.63	34.38	23.94	36.27	35.18	27.92	34.29
	HM	65.82	68.81	52.71	45.75	43.56	57.86	58.64	46.08	54.90
CoCoOp [CVPR22]	Seen	87.83	91.87	71.80	72.77	61.03	73.57	89.63	79.50	78.50
	Unseen	37.90	47.30	31.05	34.78	25.07	37.20	41.46	28.27	35.38
	HM	65.21	67.74	52.39	39.62	41.19	58.04	62.25	45.85	54.04
VPT [ECCV22]	Seen	90.68	93.12	75.70	58.82	66.30	69.94	88.56	76.73	77.48
	Unseen	35.59	44.59	31.72	26.19	26.29	28.47	37.94	23.92	31.83
	HM	60.31	65.73	50.75	35.63	40.61	48.32	58.02	46.40	50.72
PromptSRC [ICCV23]	Seen	91.21	93.20	75.60	74.27	67.23	73.67	89.73	82.65	80.95
	Unseen	40.75	50.27	32.73	35.33	27.97	40.10	43.15	28.52	37.35
	HM	66.43	70.02	53.22	42.12	45.58	58.16	63.87	48.40	55.98
DePT [CVPR24]	Seen	91.28	93.67	75.86	62.13	70.37	70.27	90.30	78.73	79.08
	Unseen	35.90	44.63	31.47	40.30	28.26	30.13	40.87	30.10	35.21
	HM	63.34	67.85	51.91	50.83	48.86	54.65	65.41	54.25	57.14
CoPL [AAAI24]	Seen	88.60	91.40	75.43	73.94	66.20	73.41	90.82	82.36	80.27
	Unseen	39.54	50.70	32.69	34.90	25.72	40.61	42.70	28.18	36.88
	HM	66.41	69.03	53.06	41.55	45.10	59.29	61.75	49.16	55.67
GalLoP [ECCV24]	Seen	88.15	91.69	74.38	74.07	67.43	73.84	90.15	82.47	80.27
	Unseen	39.33	50.55	31.67	35.83	27.22	41.39	42.67	29.41	37.26
	HM	66.16	69.05	52.70	41.90	45.26	60.02	60.86	49.71	55.71
ProText [AAAI25]	Seen	86.84	92.13	71.87	73.40	70.92	76.88	91.35	82.90	80.79
	Unseen	37.45	48.43	32.65	40.95	30.80	41.43	42.87	35.47	38.77
	HM	65.67	68.24	56.45	50.15	46.76	65.49	67.09	56.27	59.52
MMRL [CVPR25]	Seen	89.75	**95.86**	76.35	75.30	68.70	76.74	90.57	83.20	82.06
	Unseen	38.77	48.59	33.19	40.07	30.27	42.50	45.24	36.30	39.37
	HM	66.56	70.58	54.97	49.06	49.23	65.03	68.15	58.25	60.23
VMamba-B† [NeurIPS24]	Seen	85.27	92.03	80.37	73.61	58.32	80.68	81.83	79.87	78.99
	Unseen	50.93	52.20	27.78	36.95	33.76	47.95	43.06	28.55	40.15
	HM	65.57	71.18	49.28	45.10	42.26	56.15	60.63	53.76	55.49
CLIP-B/32† [ICML21]	Seen	87.48	92.15	84.51	75.52	73.79	90.94	88.97	85.34	84.83
	Unseen	60.71	59.72	35.25	50.49	41.05	45.26	46.50	39.51	47.31
	HM	72.90	76.33	59.37	55.34	60.18	69.63	65.22	66.18	65.64
SigLIP-B/16† [ICCV23]	Seen	87.85	94.02	83.50	77.06	73.52	**93.33**	**93.81**	**91.18**	86.78
	Unseen	63.65	60.16	39.93	60.15	45.70	51.01	48.21	47.57	52.05
	HM	73.92	77.95	62.25	62.54	60.39	70.08	69.19	64.02	67.69
ProME-DR (ours)	Seen	**92.07**	95.34	**85.99**	**80.35**	**80.50**	92.15	91.01	84.69	**87.76**
	Unseen	**70.68**	**77.03**	**46.56**	**76.92**	**65.87**	**72.69**	**68.78**	**65.79**	**68.04**
	HM	**78.33**	**82.81**	**65.25**	**68.16**	**64.28**	**75.47**	**71.56**	**68.61**	**71.81**

Table 1: **Seen-to-Unseen grades** performance of ProME-DR and other baselines on eight different DR datasets. 'HM' refers to Harmonic Mean (Zhou et al. 2022a), † represents the pre-trained model is fully fine-tuned (Dong et al. 2022).

et al. 2024), DePT (Zhang et al. 2024), CoPL (Goswami et al. 2024), ProText (Khattak et al. 2025), MMRL (Guo and Gu 2025); 2) fine-tuned: CLIP (Radford et al. 2021), SigLIP (Zhai et al. 2023), VMamba (Liu et al. 2024) fully fine-tuned by strategy (Dong et al. 2022); 3) single-view DR grading: SwinTransformer (Liu et al. 2021), ConvNeXt (Liu et al. 2022c), PVT (Wang et al. 2021); 4) multi-view DR grading: MVCINN (Luo et al. 2023), MVTSM (Lin et al. 2025), SMVDR (Luo et al. 2025).

Implementation Details. In all our experiments, we use a fixed image encoder and text encoder initialized with pre-trained CLIP (ViT-B/32). The Concept-Aware Emulator (CAE) is built with a two-layer ViT (Dosovitskiy et al. 2020) model with input and output dimensions of 768. The visual prompt dimension $\mathbb{R}^d$ is set to $d = 768$, and text prompt dimension $\mathbb{R}^\ell$ is set to $\ell = 512$. For the DR prompts $\mathcal{T}_{\mathtt{DR}} \in \mathbb{R}^{K\times\ell}$, we set $K = 16$, same as (Zhou et al. 2022b,a). All our experiments are trained on two 24 GB NVIDIA RTX A5000 GPUs with a batch size of 64, training epochs 10 in Stage-1, and epochs 100 in Stage-2. We set $\tau = 0.06$ in Eq. (11), $\lambda_1 = 0.5$ in Eq. (8), $\lambda_2 = 1.0$ in Eq. (12). Our starting learning rate is $2e^{-3}$ with a cosine learning rate scheduler.

Seen to Unseen DR Grade/View Generalization

As shown in Table 1, we report the accuracy of Seen and Unseen classes and their Harmonic Mean (HM), averaged over 3 runs on the above 8 DR datasets. We observe that the performance of these baselines degrades if compared to the fully fine-tuned CLIP or SigLIP for unseen classes. Actually, the visual clues of the semantics of DR samples are usually more subtle and ambiguous than most common-life objects in natural images, thus most prompt-based methods perform poorly. Regarding the average performance across these DR datasets, ProME-DR continues to maintain a substantial lead, as evidenced by a notable 4.12% (HM) improvement over the fully fine-tuned SigLIP. The results suggest that our two-stage training of matching and then emulating benefits the semantic understanding of unseen DR samples. The results of the Seen-to-Unseen DR view generalization on MFIDDR are provided in Appendix B.

Compared Methods	Source	Targets							
	MFIDDR	APTOS	DeepDR	IDRiD	MESSIDOR	RLDR	DDR	EyePACS	**Average**
DePT [CVPR24]	70.97	67.01	48.23	48.35	42.67	55.70	53.40	52.07	52.49
ProText [AAAI25]	67.80	67.95	50.13	49.01	42.35	54.66	51.72	51.34	52.45
MMRL [CVPR25]	72.03	68.25	50.67	49.43	42.10	55.57	53.69	51.45	53.22
CLIP-B/32† [ICML21]	74.63	68.30	52.32	50.70	43.97	57.23	55.28	51.86	54.37
SigLIP-B/16† [ICCV23]	75.71	69.77	50.63	52.40	46.93	60.38	56.47	55.18	55.97
ProME-DR (ours)	**87.90**	**73.96**	**55.89**	**60.75**	**52.83**	**68.58**	**70.63**	**66.49**	**64.16**

Table 2: **Cross-Dataset generalization** performance of ProME-DR with baselines on cross-dataset evaluation. Source is the MFIDDR dataset, targets are the other seven datasets, and † represents the pre-trained model that is fully fine-tuned.

Methods	Acc.	Spec.	Kappa	F1	AUC
Swin-B†	75.08	75.53	51.32	72.42	88.83
PVT-L†	75.31	80.39	57.29	73.89	90.38
ConvNeXt-B†	75.96	77.81	53.72	73.65	89.77
MVCINN	80.10	83.32	62.45	78.86	91.07
MVTSM	71.73	82.76	-	70.93	-
SMVDR	81.64	**89.92**	66.94	81.38	-
ProME-DR	**87.90**	85.29	**77.21**	**87.14**	**97.26**

Table 3: **Supervised evaluation** of ProME-DR with single-view and multi-view methods on MFIDDR using accuracy (Acc.), specificity (Spec.), Kappa, F1, and AUC. † represents the pre-trained model is fully fine-tuned.

Cross-Dataset DR Generalization

We further demonstrate that our ProME-DR has the potential to transfer knowledge beyond a single dataset. As shown in Table 2, each model is only trained on the MFIDDR dataset and directly validated on test sets of the other 7 datasets. We can observe that our ProME-DR achieves superior performance compared to others, exhibiting stronger generalizability and transferability. These results indicate that our ProME-DR can well adapt CLIP to extract generalized visual features even in the cross-dataset scenario.

Supervised Evaluation on MFIDDR

We report the results against SOTA single-/multi-view DR grading on the MFIDDR dataset in Table 3. Note that the multi-view MVTSM (Lin et al. 2025) and SMVDR (Luo et al. 2025) do not contain additional guidance (*e.g.*, lesion information). Table 3 shows that ProME-DR consistently outperforms popular counterparts in multiple metrics. These improvements are largely attributed to our two-stage training strategy (matching and then emulating), which enables ProME-DR to polish DR-specific visual clues.

Ablative Study and Analysis

We present average HM performance with different τ in Eq. (11) under the seen-to-unseen DR grades scenario. Fig. 3 (a) reveals that a smaller τ benefits training more than higher ones, which is confirmed by literature (Wu et al. 2023). Accuracy progressively increases as τ enlarges, and the improvement becomes marginal when $\tau = 0.06$, we choose $\tau = 0.06$ by default. Fig. 3 (b) reports average HM accuracy of various λ_1 (Eq. (8)) & λ_2 (Eq. (12)), we can see that:

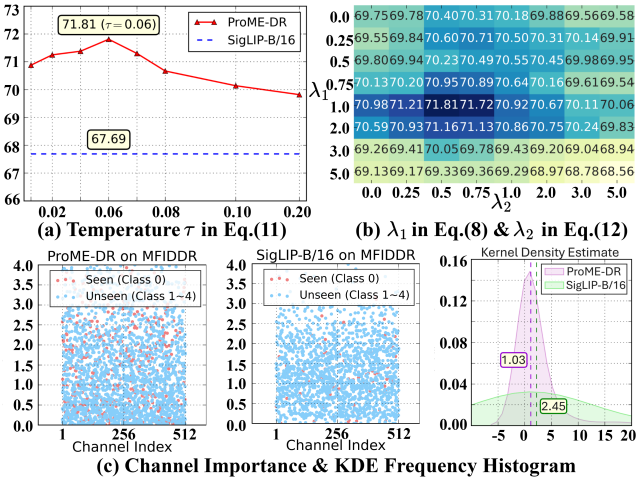


Figure 3: **Ablative study** for our proposed ProME-DR.

each loss term is indispensable (first row means without $\mathcal{L}_{cl}$, first column means without $\mathcal{L}_{dis}$) and combined together can reach better performance ($\lambda_1 = 0.5, \lambda_2 = 1.0$).

Fig. 3 (c) draws channel importance (CI) (Zhang et al. 2024) distributions on MFIDDR by ProME-DR and SigLIP model, we calculate the CI (remove NaN) of seen and unseen classes, and plot the frequency histogram of seen-CI / unseen-CI. We observe that CI distributions of seen and unseen classes obtained by our ProME-DR show better consistency compared to SigLIP. The histogram further shows that most of the SigLIP's learned features are occupied by specific seen knowledge, resulting in poor generalization to unseen samples, while ProME-DR alleviates this channel bias.

Conclusion

In this paper, we propose ProME-DR, a two-stage prompt-driven framework for zero-shot DR grading, which would enable CLIP to extract generalized knowledge for recognizing unseen DR samples. It disentangles the training into two stages: PML effectively captures class-relevant and inter-view consistency knowledge; PEL dynamically produces context prompt units via a concept-aware emulator to adapt to previously unseen samples. The effectiveness of ProME-DR has been validated on various datasets and tasks.

Acknowledgments

This work was supported by the scholarship from the China Scholarship Council (CSC) while the first author pursued his PhD degree at the University of Wollongong. We thank Prof. Guansong Pang (SMU) and Prof. Jun Shen (UOW) for their valuable feedback and discussions on this paper. This work was also partially supported by Australian Research Council Linkage Project LP210300009 and LP230100083.

References

Abràmoff, M. D.; Folk, J. C.; Han, D. P.; et al. 2013. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmology*, 131(3): 351–357.

Atwany, M. Z.; Sahyoun, A. H.; and Yaqub, M. 2022. Deep learning techniques for diabetic retinopathy classification: A survey. *IEEE Access*, 10: 28642–28655.

Bafghi, R. A.; Harilal, N.; Monteleoni, C.; and Raissi, M. 2024. Parameter Efficient Fine-tuning of Self-supervised ViTs without Catastrophic Forgetting. In *CVPR*, 3679–3684.

Bai, A.; Yeh, C.-K.; Hsieh, C.-J.; and Taly, A. 2025. An Efficient Rehearsal Scheme for Catastrophic Forgetting Mitigation during Multi-stage Fine-tuning. In *NAACL*, 2557–2569.

Che, H.; Cheng, Y.; Jin, H.; and Chen, H. 2023. Towards generalizable diabetic retinopathy grading in unseen domains. In *MICCAI*, 430–440. Springer.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Dai, L.; Sheng, B.; Chen, T.; Wu, Q.; et al. 2024. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*, 30(2): 584–594.

Dai, L.; Wu, L.; Li, H.; Cai, C.; Wu, Q.; Kong, H.; Liu, R.; Wang, X.; Hou, X.; Liu, Y.; et al. 2021. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature Communications*, 12(1): 3242.

Dai, W.; Mou, C.; Wu, J.; et al. 2023. Diabetic retinopathy detection with enhanced vision transformers: The twins-pcpvt solution. In *ICETCI*, 403–407. IEEE.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; et al. 2022. CLIP Itself is a Strong Fine-tuner: Achieving 85.7% and 88.0% Top-1 Accuracy with ViT-B and ViT-L on ImageNet. *arXiv preprint arXiv:2212.06138*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fang, A.; Ilharco, G.; Wortsman, M.; Wan, Y.; et al. 2022. Data determines distributional robustness in contrastive language image pre-training (clip). In *ICML*, 6216–6234. PMLR.

Fang, K.; Song, J.; Gao, L.; Zeng, P.; Cheng, Z.-Q.; Li, X.; and Shen, H. T. 2024. ProS: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *CVPR*, 17292–17301.

Goswami, K.; Karanam, S.; Udhayanan, P.; et al. 2024. Copl: Contextual prompt learning for vision-language understanding. In *AAAI*, 18090–18098.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 21271–21284.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410.

Guo, Y.; and Gu, X. 2025. MMRL: Multi-modal representation learning for vision-language models. In *CVPR*, 25015–25025.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; et al. 2022. A survey on vision transformer. *IEEE TPAMI*, 45(1): 87–110.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Heng, L.; Comyn, O.; Peto, T.; Tadros, C.; et al. 2013. Diabetic retinopathy: pathogenesis, clinical grading, management and future developments. *Diabetic Medicine*, 30(6): 640–650.

Hu, J.; Chen, R.; Lu, Y.; Dou, X.; Ye, B.; Cai, Z.; Pu, Z.; and Mou, L. 2019. Single-field non-mydriatic fundus photography for diabetic retinopathy screening: a systematic review and meta-analysis. *Ophthalmic Research*, 62(2): 61–67.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; et al. 2022. Visual prompt tuning. In *ECCV*, 709–727. Springer.

Kalyani, G.; Janakiramaiah, B.; Karuna, A.; et al. 2023. Diabetic retinopathy detection and classification using capsule networks. *Complex & Intelligent Systems*, 9(3): 2651–2664.

Karthik; Maggie; and Dane, S. 2019. APTOS 2019 Blindness Detection. https://kaggle.com/competitions/aptos2019-blindness-detection. Kaggle. Accessed May 18, 2025.

Khattak, M. U.; Naeem, M. F.; Naseer, M.; Van Gool, L.; and Tombari, F. 2025. Learning to prompt with text only supervision for vision-language models. In *AAAI*, 4230–4238.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; et al. 2023. Self-regulating Prompts: Foundational Model Adaptation without Forgetting. In *ICCV*, 15190–15200.

Kumar, R.; Kumbharkar, P.; Vanam, S.; and Sharma, S. 2024. Medical images classification using deep learning: a survey. *Multimedia Tools and Applications*, 83(7): 19683–19728.

Lafon, M.; Ramzi, E.; Rambour, C.; Audebert, N.; and Thome, N. 2024. Gallop: Learning global and local prompts for vision-language models. In *ECCV*, 264–282. Springer.

Lai, H.; Yao, Q.; Jiang, Z.; Wang, R.; et al. 2024. Carzero: Cross-attention alignment for radiology zero-shot classification. In *CVPR*, 11137–11146.

Li, T.; Gao, Y.; Wang, K.; Guo, S.; et al. 2019. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501: 511–522.

Li, Z.; Li, X.; Fu, X.; et al. 2024. PromptKD: Unsupervised Prompt Distillation for Vision-Language Models. In *CVPR*, 26617–26626.

Lin, Y.; Dou, X.; Luo, X.; Wu, Z.; et al. 2025. Multi-view diabetic retinopathy grading via cross-view spatial alignment and adaptive vessel reinforcing. *Pattern Recognition*, 164: 111487.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; et al. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, R.; Wang, X.; Wu, Q.; Dai, L.; et al. 2022a. DeepDRiD: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge. *Patterns*, 100512.

Liu, R.; Wang, X.; Wu, Q.; Dai, L.; et al. 2022b. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6): 100512.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. In *NeurIPS*, 103031–103063.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; et al. 2022c. A convnet for the 2020s. In *CVPR*, 11976–11986.

Luo, X.; Liu, C.; Wong, W.; Wen, J.; et al. 2023. MVCINN: multi-view diabetic retinopathy detection using a deep cross-interaction neural network. In *AAAI*, volume 37, 8993–9001.

Luo, X.; Pu, Z.; Xu, Y.; Wong, W. K.; et al. 2021. MVDR-Net: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. *Pattern Recognition*, 120: 108104.

Luo, X.; Xu, Q.; Wang, Z.; Huang, C.; et al. 2024. A Lesion-Fusion Neural Network for Multi-View Diabetic Retinopathy Grading. *JBHI*, 29(5): 3184–3193.

Luo, X.; Xu, Q.; Wu, H.; Liu, C.; et al. 2025. Like an Ophthalmologist: Dynamic Selection Driven Multi-View Learning for Diabetic Retinopathy Grading. In *AAAI*, volume 39, 19224–19232.

Martin, S.; Huang, Y.; Shakeri, F.; et al. 2024. Transductive zero-shot and few-shot clip. In *CVPR*, 28816–28826.

Min, B.; Ross, H.; Sulem, E.; et al. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40.

Padhy, S. K.; Takkar, B.; Chawla, R.; and Kumar, A. 2019. Artificial intelligence in diabetic retinopathy: A natural step to the future. *Ophthalmology*, 67(7): 1004–1009.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; et al. 2018. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, 3(3): 25.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; et al. 2022. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183: 109–119.

Suzuki, K. 2017. Overview of deep learning in medical imaging. *Radiological Physics and Technology*, 10(3): 257–273.

Teo, Z. L.; Tham, Y.-C.; Yu, M.; Chee, M. L.; et al. 2021. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*, 128(11): 1580–1591.

Ursin, F.; Timmermann, C.; Orzechowski, M.; and Steger, F. 2021. Diagnosing diabetic retinopathy with artificial intelligence: What information should be included to ensure ethical informed consent? *Frontiers in Medicine*, 8: 695217.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; et al. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 568–578.

Wei, Q.; Li, X.; Yu, W.; Zhang, X.; et al. 2021. Learn to segment retinal lesions and beyond. In *ICPR*, 7403–7410. IEEE.

Wu, J.; Chen, J.; Wu, J.; Shi, W.; Wang, X.; and He, X. 2023. Understanding contrastive learning via distributionally robust optimization. In *NeurIPS*, volume 36, 23297–23320.

Wu, J.; Yu, T.; Wang, R.; Song, Z.; et al. 2024. InfoPrompt: Information-Theoretic Soft Prompt Tuning for Natural Language Understanding. In *NeurIPS*, 61060–61084.

Xia, Y.; Kim, J.; Chen, Y.; Ye, H.; et al. 2024. Understanding the performance and estimating the cost of llm fine-tuning. In *IISWC*, 210–223. IEEE.

Xing, J.; Liu, J.; Wang, J.; Sun, L.; et al. 2024. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers and Graphics*, 119: 103885.

You, H.; Zhou, L.; Xiao, B.; Codella, N.; et al. 2022. Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, 69–87.

Yu, Z.; Dong, Z.; Yu, C.; Yang, K.; et al. 2025. A review on multi-view learning. *Frontiers of Computer Science*, 19(7): 197334.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *ICCV*, 11975–11986.

Zhang, J.; Wu, S.; Gao, L.; Shen, H. T.; and Song, J. 2024. Dept: Decoupled prompt tuning. In *CVPR*, 12924–12933.

Zhang, W.-f.; Li, D.-h.; Wei, Q.-j.; Ding, D.-y.; Meng, L.-h.; Wang, Y.-l.; Zhao, X.-y.; and Chen, Y.-x. 2022. The validation of deep learning-based grading model for diabetic retinopathy. *Frontiers in Medicine*, 9: 839088.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *CVPR*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *IJCV*, 130(9): 2337–2348.