CLIP-DR: Textual Knowledge-Guided Diabetic Retinopathy Grading with Ranking-aware Prompting

Qinkai Yu[1], Jianyang Xie[2], Anh Nguyen[3], He Zhao[2], Jiong Zhang[4], Huazhu Fu[5], Yitian Zhao[4], Yalin Zheng[2], and Yanda Meng[1(✉)]

[1] Computer Science Department, University of Exeter, Exeter, UK
[2] Eye and Vision Sciences Department, University of Liverpool, Liverpool, UK
[3] Computer Science Department, University of Liverpool, Liverpool, UK
[4] Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
[5] Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore.
Y.M.Meng@exeter.ac.uk

Abstract. Diabetic retinopathy (DR) is a complication of diabetes and usually takes decades to reach sight-threatening levels. Accurate and robust detection of DR severity is critical for the timely management and treatment of diabetes. However, most current DR grading methods suffer from insufficient robustness to data variability (*e.g.* colour fundus images), posing a significant difficulty for accurate and robust grading. In this work, we propose a novel DR grading framework CLIP-DR based on three observations: 1) Recent pre-trained visual language models, such as CLIP, showcase a notable capacity for generalisation across various downstream tasks, serving as effective baseline models. 2) The grading of image-text pairs for DR often adheres to a discernible natural sequence, yet most existing DR grading methods have primarily overlooked this aspect. 3) A long-tailed distribution among DR severity levels complicates the grading process. This work proposes a novel ranking-aware prompting strategy to help the CLIP model exploit the ordinal information. Specifically, we sequentially design learnable prompts between neighbouring text-image pairs in two different ranking directions. Additionally, we introduce a Similarity Matrix Smooth module into the structure of CLIP to balance the class distribution. Finally, we perform extensive comparisons with several state-of-the-art methods on the GDRBench benchmark, demonstrating our CLIP-DR's robustness and superior performance. The implementation code is available [1].

Keywords: Vision language model · Diabetic retinopathy grading · Rank-aware prompt learning

[1] https://github.com/Qinkaiyu/CLIP-DR

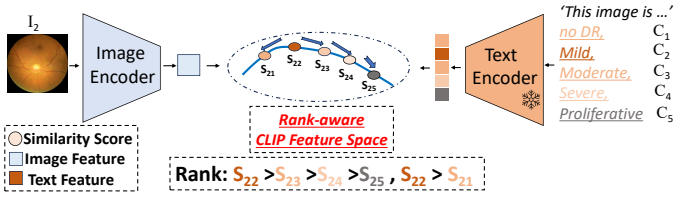

Fig. 1: **An example of learnable rank-aware prompting with image class 'Mild' for a given Image I_2.** $[C_1, ..., C_5]$ represent 5 different DR grading classes. The similarity score is obtained by the inner product of the image feature and text feature. Designing learnable rank-aware prompts that satisfy the following two inequalities enables the model to learn natural order information.

1 Introduction

Diabetic retinopathy (DR) is a main complication of diabetes, typically progressing over several years before reaching levels that threaten vision. The disease process can present a wide range of severity grades and change over time. These grades of severity are often categorised into different classes (*e.g.*, normal, mild, moderate, severe [1]), and the variations among these classes are often difficult to discern. Stylistic appearance variability in the different data sources of colour fundus images further complicates this challenge [2]. A growing number of deep learning studies researched the DR grading tasks with colour fundus images [3,4,5,6,7,8]. However, most of these previous studies attribute the poor performance of classification models in DR grading to the diversity of diagnostic patterns, data imbalance, and large differences in individual appearance styles. Recently, the Contrastive Language-Image Pre-Training (CLIP) model [9] has gained significant attention for its strong generalisation capability in various downstream vision tasks, particularly achieving high classification accuracy for unbalanced or stylistically diverse data. Many paradigms of CLIP in the medical domain have also shown great potential. For instance, CLIP performance was enhanced by prompt strategies for combining medical knowledge [10]. In this work, we propose a novel framework (an example is shown in Fig. 1), CLIP-DR, that takes advantage of CLIP's robust and sufficient feature learning ability.

Most DR grading studies [11,12,13,14,1] assumed that DR labels were independent. The different groups' ground truth was converted to a one-hot format during training. However, the fact that the grade of DR follows an underlying natural order was ignored. The commonly adopted loss function of Cross-Entropy by previous methods [15] resulted in the same penalty regardless of the category to which the sample was misclassified. Typically, for the tasks with a natural order such as age estimation [16], the regular approach is to treat classification as a metric regression problem [17], minimising the absolute/squared error loss (*i.e.*, MAE/MSE). However, by treating discrete data as continuous values, the regression models are prone to ambiguity in the boundaries between different classes, making it difficult to distinguish between neighbouring classes [18]. Therefore,

it is not appropriate to approach DR grading purely from a classification or regression point of view.

The challenges above lead to the question: **Can we think out of the box and rationalise the fact that DR conforms to the natural order to improve grading performance?** We propose a new CLIP-based framework (CLIP-DR, pipeline shown in Fig. 2) treating DR grading as an image-text matching problem. Additionally, we propose a ranking-aware prompting strategy to fine-tune the CLIP image encoder. This enables CLIP-DR to learn the associations of natural ordering information under a classification task. Specifically, we minimise the Kullback-Leibler (KL) divergence as the primary loss function. The ranking loss function is proposed as ranking-aware prompting, empowering the image encoder to grasp natural ordering information. Its purpose is to ensure that the ranking aligns consistently with the inherent order of DR. We further improve the grading performance by introducing a Similarity Matrix Smooth (SMS) module into the framework to minimise the impacts due to data imbalance/long-tailed distribution. We compare the performance of CLIP-DR with other state-of-the-art methods on benchmark GDRBench [14]. Experimental results demonstrate CLIP-DR's effectiveness and superior performance.

2 Methods

2.1 Problem Formulation

Given an input colour fundus image I_i, and its corresponding class C_i, a pair of them $\{I_i, C_i\}$ is an element of DR grading dataset $\mathcal{D}$. Suppose there is K number of different classes; the set $\mathcal{T} = \{t_j\}_{j=1}^{K}$ can be defined as a set of text embeddings of classes C. We also define x_i as the image embeddings of I_i from the image encoder and set $\mathcal{X} = \{x_i\}_{i=1}^{M}$. Where M is the number of images. With CLIP, text and images can be mapped to the same dimension in our framework, such as $x_i, t_j \in \mathbb{R}^{1\times 1024}$. The similarity matrix $\mathcal{S}$ between set $\mathcal{X}$ and $\mathcal{T}$ is calculated via inner product, where $\mathcal{S} = [s_{i,j}]_{M\times K} \in \mathbb{R}^{M\times K}$, and $s_{i,j} = x_i \cdot t_j^T$. Then for a given sample $\{I_i, C_i\}$, our aim is to learn a mapping $f_\theta : I_i \to C_i$, where f is a deep neural network with parameters θ.

2.2 SMS for Imbalanced Regression

This section presents the Similarity Matrix Smooth (*SMS*) module. We integrate SMS into CLIP-DR by inserting a feature calibration layer after the similarity matrix. In essence, SMS is a mapping $g : \mathbb{R}^{M\times K} \to \mathbb{R}^{M\times K} : \mathcal{S} \to \tilde{\mathcal{S}}$ designed to smooth the original similarity matrix. This smoothing process is inspired by [19] and aims to mitigate the impact of data imbalance, with the resulting smoothed statistic reflecting the ordinal relationship among neighbouring targets. Our SMS addresses data imbalance by smoothing similarity vectors, while [19] targets

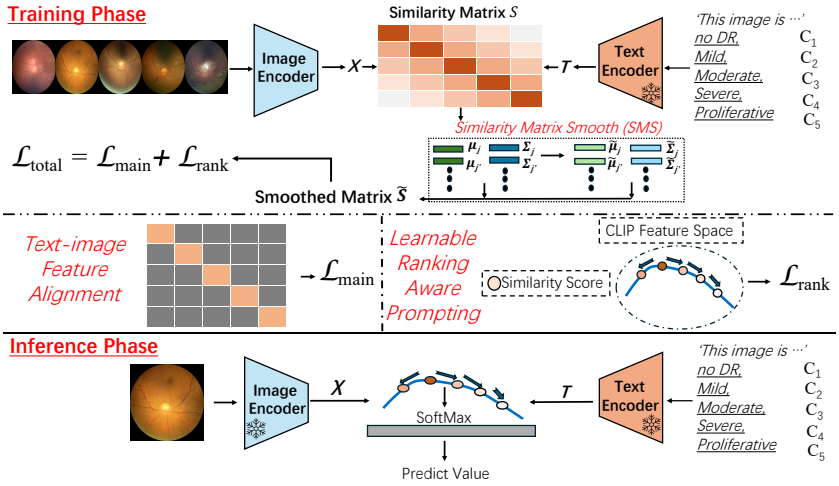

Fig. 2: Overview of the proposed CLIP-DR framework for training and inference. Images are processed through an image encoder to extract image features X. The corresponding text labels are fed into the text encoder, generating text embeddings for labels T. The similarity matrix $\mathcal{S}$ is obtained through the inner product. Finally, the SMS module converts $\mathcal{S}$ into calibration features $\tilde{\mathcal{S}}$ with the same dimensions. The learnable rank-aware prompt strategy is implemented explicitly by $\mathcal{L}_{\text{rank}}$, which uses ranking information independently in the left and right directions, and $\mathcal{L}_{\text{main}}$ follows the practice of CLIP [9].

entropy. We define $s_{:,j} \in \mathbb{R}^{1\times K}$ as the row vector that the true text embedding is t_j. By estimating the statistics of $s_{:,j}$, we can easily obtain the mean and variance. Then we define μ_j as the mean of all row vectors $s_{:,j}$ and Σ_j as the variance. After employing a symmetric kernel $\mathcal{K}$ to smooth the distribution of μ_j and Σ_j, they can be defined as:

$$\tilde{\mu}_j = \sum_{j'\neq j}^{K} \mathcal{K}(Y_{:,j}, Y_{:,j'})\mu_{j'}, \quad \tilde{\Sigma}_j = \sum_{j'\neq j}^{K} \mathcal{K}(Y_{:,j}, Y_{:,j'})\Sigma_{j'} \tag{1}$$

Then we calibrate the similarity vector $s_{:,j}$ into $\tilde{s}_{:,j}$, such as:

$$\tilde{s}_{:,j} = \tilde{\Sigma}_j^{\frac{1}{2}} \tilde{\mu}_j^{\frac{1}{2}} (s_{:,j} - \mu_j) + \tilde{\mu}_j \tag{2}$$

$\tilde{\mu}_j$ and $\tilde{\Sigma}_j$ updated across different epochs but fixed within each training epoch. Finally, stacking $\tilde{s}_{:,j}$ by rows yields $\tilde{\mathcal{S}}$. Such a smoothed matrix $\tilde{\mathcal{S}}$ can calibrate potentially biased estimates of similarity matrix distributions, especially for classes with few samples, thus mitigating the impact of data imbalance for prompt learning.

2.3 Prompting Rank-aware Gradient

The core idea of rank-aware prompting is to exploit the natural order of DR colour fundus images, thus improving the grading accuracy. The motivation comes from a reasonable explanation: misdiagnosis of the highest-ranked patient as having no disease is more severe than misdiagnosis of the highest-ranked patient as having an intermediate disease. The details of rank-aware gradient prompting are elaborated below.

Main Loss. Our main loss consists of image-to-text loss and text-to-image loss. Recall that if t_j is the text embeddings corresponding to I_i, then $s_{i,j}$ is the similarity score of this correct matching. $\mathcal{S} \in \mathbb{R}^{M \times K}$ is a text-to-image pair matrix, while $\mathcal{S}^T \in \mathbb{R}^{K \times M}$ is a image-to-text pair matrix. The SMS module converts $\mathcal{S}$ to $\tilde{\mathcal{S}}$. Thus the corresponding label can be expressed as Y for text-to-image pair in Eq. 3 and Y^T for text-to-image pair:

$$Y = \mathbb{I}(\tilde{\mathcal{S}}) = \begin{cases} 1, \text{ correct matching} \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

Given that $K \leq M$ and $\tilde{\mathcal{S}}$ is not square, the image-to-text label (Y^T) could potentially yield zero or multiple hits. Consequently, treating similarity score learning as a 1-in-N classification problem with cross-entropy loss, akin to the original CLIP framework [20], would be unsuitable. Thus, inspired by previous methods [20,21], we adopt the KL divergence loss. Specifically, we construct the new label matrix Y' using the non-zero columns of the normalised transpose of Y^T. $\tilde{\mathcal{S}}^T$ is also applied via the softmax layer to obtain the normalised matrix ${\tilde{\mathcal{S}}'}^T$. Therefore, image-to-text loss and text-to-image loss can be expressed as:

$$\text{image-to-text: } \frac{1}{K}\sum_{i=1}^{K} KL(Y'_{i,\cdot} \| {\tilde{\mathcal{S}}'}^T_{i,\cdot}), \quad \text{text-to-image: } \frac{1}{M}\sum_{i=1}^{M} KL(Y_{i,\cdot} \| {\tilde{\mathcal{S}}'}_{i,\cdot}) \tag{4}$$

Note that the normalised Y is equal to itself since there is only one hit in each row and $\tilde{\mathcal{S}}'$ is the normalised matrix by $\tilde{\mathcal{S}}$. At the end, the main loss is defined in Eq. 5 as:

$$\mathcal{L}_{\text{main}} = \frac{1}{2}[\frac{1}{K}\sum_{i=1}^{K} KL(Y'_{i,\cdot} \| {\tilde{\mathcal{S}}'}^T_{i,\cdot}) + \frac{1}{M}\sum_{i=1}^{M} KL(Y_{i,\cdot} \| {\tilde{\mathcal{S}}'}_{i,\cdot})] \tag{5}$$

Rank Loss. For a given image I_i, t_j is its corresponding text embeddings and $s_{i,j}$ is the similarity score of this correct matching. The DR coloured fundus of the overall dataset has a natural grade of information, although very few patients might experience, for example, a direct transition from mild or moderate DR to proliferative DR, without experiencing severe DR. We aim to prompt the rank information of DR grading by ensuring that:

$$\begin{cases} \tilde{s}_{i,j} > \tilde{s}_{i,j+1} > ... > \tilde{s}_{i,K} \\ \tilde{s}_{i,j} > \tilde{s}_{i,j-1} > ... > \tilde{s}_{i,1} \end{cases} \tag{6}$$

To this end, we propose a novel rank-loss that performs binary classification for each neighbouring class in two directions (left and rightward). For example, for a neighbouring class pair $(\tilde{s}_{i,j'}, \tilde{s}_{i,j'+1})$, we design the loss function by minimising the gap between this pair and the label (1, 0). Thus, the rank loss function ($\mathcal{L}_{\text{rank}}$) is defined in Eq. 7 as:

$$\mathcal{L}_{\text{rank}} = -\frac{1}{M}\sum_{i=1}^{M}(\mathcal{L}^{i}_{\text{rightward}} + \mathcal{L}^{i}_{\text{leftward}}) \tag{7}$$

where $\mathcal{L}^{i}_{\text{rightward}}$ and $\mathcal{L}^{i}_{\text{leftward}}$ are the learnable prompt approach designed to fit the goals in Eq. 6. They are defined as:

$$\mathcal{L}^{i}_{\text{rightward}} = -\sum_{j'=j}^{k-1}\log\frac{\exp(\tilde{s}_{i,j'}/\tau)}{\exp(\tilde{s}_{i,j'}/\tau)+\exp(\tilde{s}_{i,j'+1}/\tau)}, \tag{8}$$

$$\mathcal{L}^{i}_{\text{leftward}} = -\sum_{j'=2}^{j}\log\frac{\exp(\tilde{s}_{i,j'}/\tau)}{\exp(\tilde{s}_{i,j'}/\tau)+\exp(\tilde{s}_{i,j'-1}/\tau)}, \tag{9}$$

where we set label of $\tilde{s}_{i,j'}$, $\tilde{s}_{i,j'+1}$ and $\tilde{s}_{i,j'-1}$ is 1, 0, 0, respectively. The τ was set to 1 during the experiment. It is worth noting that although the proposed $\mathcal{L}_{\text{rank}}$ is similar to the kappa loss [22] (the kappa loss is usually used for classification of ordered labels), unlike fixed weights in kappa loss which are linear or quadratic, the weights in $\mathcal{L}_{\text{rank}}$ are in a self-adaptive non-linear fashion, making the $\mathcal{L}_{\text{rank}}$ applicable to the learnable prompting text-image alignment in clip feature space. In the end, the total loss is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda\mathcal{L}_{\text{rank}}$, where λ is a hyper-parameter weighting the loss term $\mathcal{L}_{\text{rank}}$, which is set as 1 by default.

3 Datasets and Implementation Details

We follow the same experiment setting of GDRBench [14] involving two generalization ability evaluation settings and eight popular public datasets.

Evaluation Settings. The initial evaluation follows the classic leave-one-domain-out protocol (DG test), where one domain is withheld for evaluation while models are trained on the remaining domains. The DG test encompasses six datasets: DeepDR [23], Messidor [24], IDRID [25], APTOS [26], FGADR [27], and RLDR [28]. Another evaluation scenario is the extreme single-domain generalization setting (ESDG test), which adopts a train-on-single-domain protocol using the aforementioned datasets. We adopted two essential metrics to evaluate the performance: the area under the ROC curve (AUC) and macro F1-score (F1). We used **bold** and underline to indicate the first and the second-highest scores in each sub-dataset test performance.

Implementation Details. All experiments are conducted on an Nvidia 4090 GPU. CLIP [9], OrdinalCLIP [21], and our CLIP-DR used pre-trained ResNet50 [29] as the image encoder backbone and text encoder is a pre-trained Transformer. The initial prompt was *"This image is* {*label*}*"*. The number of training epochs is set to 100.

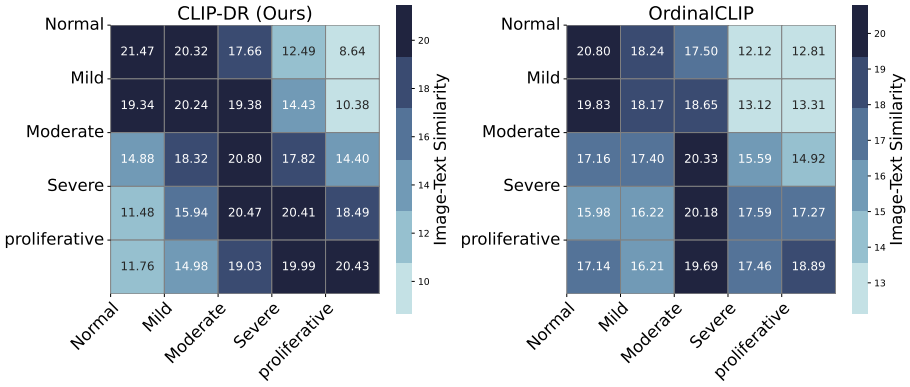

Fig. 3: The image-text similarity matrix obtained by the inner product of image feature and text feature: $X \cdot T^t$. This matrix is an intuitive representation of the rank between different image-text pairs. The X-axis represents five different text labels, and the Y-axis represents real images of different classes. We average the results of the six sub-datasets and present them in this figure for our CLIP-DR and OrdinalCLIP [21]. CLIP-DR can learn rank-aware text-image features.

4 Results

As shown in Table 1, We compare the proposed method with the following state-of-the-art methods: DRGen [3], Mixup [4], MixStyle [5], GREEN [6], CABNet [7], DDAIG [8], ATS [11], Fishr [12], MDLT [13], GDRNet [14], CLIP [9] and OrdinalCLIP [9]. Table 1 shows the results of the DG test setting, where the target row indicates the test set. Our CLIP-DR achieved the **best** performance of the F1 score on four sub-tests datasets and the second-best F1 score performance on the other two sub-tests datasets. On average, our CLIP-DR achieved the **best** performance of *F1* and *AUC* across all six datasets. Our model outperformed GDRNet[14] (SOTA) with a statistically significant p-value of 0.02, where ours achieved the best F1 or AUC. Note that Messidor [24] and RLDR [28] contain excessive differences in colour styles from the other datasets [14]. GDRNet [14] specially designed strong data enhancement techniques for finite-domain transformations to address such a challenge and thus achieved good performance. Notably, OrdinalCLIP [21] is closely related work that exploits the ordinal information when tackling regression tasks. OrdinalCLIP derives ordinal ranking embeddings $R \in \mathbb{R}^{K\times 1024}$ from a set of basic ranking embeddings $R' \in \mathbb{R}^{K'\times 1024}$ by interpolation, and the base ranking embedding length K' needs to be much smaller than the ordinal ranking embeddings length K (*e.g.*, $K' \ll K$). However, there are only five classes in DR grading, so finding such a base ranking embedding length is hard. Both our approach and OrdinalCLIP [21] aim to learn ranked feature spaces in CLIP [9]. While OrdinalCLIP [21] uses linear interpolation, we make DR fundus order learnable via text-image pairs. Compared to CLIP [9] and OrdinalCLIP [21], our CLIP-DR significantly

Table 1: Comparison with state-of-the-art approaches under the DG test.

Target	APTOS		DeepDR		FGADR		IDRID		Messidor		RLDR		Average	
Metrics	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
DRGen [3]	40.2	79.9	34.1	83.0	24.7	69.4	37.4	**84.7**	40.5	79.0	37.0	79.5	37.3	79.3
Mixup [4]	43.2	75.3	25.2	75.3	32.3	66.7	27.6	78.8	32.6	76.7	37.7	76.9	33.1	75.0
MixStyle [5]	39.9	79.0	27.9	76.9	22.7	71.2	39.2	83.0	36.5	75.2	31.4	75.5	32.9	76.8
GREEN [6]	38.9	75.1	24.9	76.4	31.5	69.5	32.2	79.9	36.8	75.8	34.4	74.8	33.1	75.3
CABNet [7]	39.4	75.8	31.8	75.2	34.8	73.2	37.3	79.2	34.1	74.2	35.6	75.8	35.5	75.6
DDAIG [8]	41.0	78.0	32.2	75.6	33.8	73.6	27.0	82.1	35.3	76.6	27.7	75.6	32.8	76.9
ATS [11]	38.3	77.1	31.6	79.4	33.4	74.7	34.9	83.0	35.8	77.2	34.9	76.5	34.8	78.0
Fishr [12]	43.4	79.2	34.4	81.1	34.4	73.3	27.6	82.7	41.1	76.4	34.7	77.4	35.9	78.4
MDLT [13]	41.5	77.3	36.2	80.0	29.0	74.1	35.4	81.5	36.9	75.4	35.0	75.7	35.7	77.3
GDRNet [14]	46.0	79.9	45.3	84.7	39.4	**80.8**	35.9	84.0	**50.9**	**83.2**	**43.5**	**82.9**	43.5	82.6
CLIP [9]	44.3	76.1	42.0	83.6	41.1	78.6	34.8	83.0	39.6	76.7	38.8	75.9	40.1	78.9
OrdinalCLIP [21]	45.7	77.6	43.3	85.1	37.9	79.3	36.2	80.4	41.8	77.7	39.5	76.6	40.7	79.4
CLIP-DR(Ours)	**46.3**	**83.3**	**45.8**	**89.9**	**48.0**	80.7	**41.9**	84.5	47.3	79.0	41.0	78.9	**45.5**	**82.7**

improves all sub-test datasets' grading performance. A heat map of the similarity matrix of OrdinalCLIP [21] and our CLIP-DR averaged over the 6 subtests is shown in Fig. 3. It shows that our CLIP-DR satisfies the objectives in Eq. 6, where a rank-aware text-image feature space is learned. While OrdinalCLIP [21] cannot reflect such rank information. To be more explicit about the effectiveness of our approach compared to OrdinalCLIP, we give the class activation map for CLIP-DR and OrdinalCLIP in the appendix. We also provide the AUC performance for each class under the DG test setting in the appendix for comprehensive experimental results. When lacking training examples, the previously widely used "pretraining-finetuning" paradigm would fail to fine-tune the entire CLIP backbone [30]. ESDG experiment setting with little training data cannot exploit the benefit of pre-trained CLIP. The intuition of this work is to boost the CLIP performance through prompt learning by exploiting the ranking information of DR images with relatively sufficient data. However, we still experimented with ESDG test setting for comprehensive experimental results; the comparison results can be found in the appendix.

Ablation study of proposed components. To assess the effectiveness of CLIP-DR, we conducted a thorough ablation study within the DG test setting and illustrated the AUC scores attained by various models in Table 2. We evaluated the performance of the model by removing the $\mathcal{L}_{\text{main}}$, the $\mathcal{L}_{\text{rank}}$, and the SMS module, individually, and remaining the rest of the model structure the same. Notably, removing the $\mathcal{L}_{\text{rank}}$ has the most significant impact on AUC (5.5 %, from an average of 82.7 to 78.4). $\mathcal{L}_{\text{main}}$, $\mathcal{L}_{\text{rank}}$, and SMS module all contribute to the improvement of the DR grading effect, such as 1.3 %, 5.5% and 2 % of AUC score. This experiment demonstrates the importance of natural order information and the effectiveness of our proposed CLIP-DR.

Table 2: Ablation studies under the DG test.

Target	APTOS		DeepDR		FGADR		IDRID		Messidor		RLDR		Average	
Metrics	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
w/o $\mathcal{L}_{\text{main}}$	44.0	82.2	40.5	86.5	33.4	80.5	37.4	83.8	45.7	78.7	38.1	78.3	39.85	81.6
w/o $\mathcal{L}_{\text{rank}}$	45.7	75.7	43.0	84.3	37.4	77.8	36.9	79.5	42.0	76.9	39.4	76.2	40.88	78.4
w/o SMS	45.5	80.9	44.9	88.2	45.6	79.9	40.0	82.9	46.4	78.1	40.3	77.0	43.78	81.1
CLIP	44.3	76.1	42.0	83.6	41.1	78.6	34.8	83.0	39.6	76.7	38.8	75.9	40.1	78.9
OrdinalCLIP	45.7	77.6	43.3	85.1	37.9	79.3	36.2	80.4	41.8	77.7	39.5	76.6	40.7	79.4
CLIP-DR	**46.3**	**83.3**	**45.8**	**89.9**	**48.0**	**80.7**	**41.9**	**84.5**	**47.3**	**79.0**	**41.0**	**78.9**	**45.5**	**82.7**

5 Conclusion

We propose a novel framework for DR grading with colour fundus images. It harnesses a ranking-aware prompting strategy to boost the vision-language model's performance by exploiting the natural ordinal information of DR image-text pairs. Extensive experiments have demonstrated that our CLIP-DR achieves state-of-the-art DR grading performance on an average of six large-scale datasets of the generalizable diabetic retinopathy grading benchmark.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. John H Kempen, Benita J O'Colmain, M Cristina Leske, Steven M Haffner, Ronald Klein, Scot E Moss, Hugh R Taylor, Richard F Hamman, et al. The prevalence of diabetic retinopathy among adults in the united states. *Archives of ophthalmology (Chicago, Ill.: 1960)*, 122(4):552–563, 2004.
2. Matthew D Li, Ken Chang, Ben Bearce, Connie Y Chang, Ambrose J Huang, J Peter Campbell, James M Brown, Praveer Singh, Katharina V Hoebel, Deniz Erdoğmuş, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. *NPJ digital medicine*, 3(1):48, 2020.
3. Mohammad Atwany and Mohammad Yaqub. Drgen: Domain generalization in diabetic retinopathy classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 635–644. Springer, 2022.
4. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
5. Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.
6. Shaoteng Liu, Lijun Gong, Kai Ma, and Yefeng Zheng. Green: a graph residual re-ranking network for grading diabetic retinopathy. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 585–594. Springer, 2020.

7. Along He, Tao Li, Ning Li, Kai Wang, and Huazhu Fu. Cabnet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1):143–153, 2020.
8. Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13025–13032, 2020.
9. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
10. Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01, 2023.
11. Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. *Advances in Neural Information Processing Systems*, 34:19448–19460, 2021.
12. Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.
13. Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision*, pages 57–75. Springer, 2022.
14. Haoxuan Che, Yuhan Cheng, Haibo Jin, and Hao Chen. Towards generalizable diabetic retinopathy grading in unseen domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 430–440. Springer, 2023.
15. Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.
16. Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016.
17. Yun Fu and Thomas S Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008.
18. Jinhong Wang, Yi Cheng, Jintai Chen, TingTing Chen, Danny Chen, and Jian Wu. Ord2seq: Regarding ordinal regression as label sequence prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5875, 2023.
19. Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pages 11842–11851. PMLR, 2021.
20. Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.
21. Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. *Advances in Neural Information Processing Systems*, 35:35313–35325, 2022.
22. Jordi de La Torre, Domenec Puig, and Aida Valls. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, 105:144–154, 2018.

23. Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019.
24. Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.
25. Prasanna Porwal, Samiksha Pachade, Manesh Kokare, Girish Deshmukh, Jaemin Son, Woong Bae, Lihong Liu, Jianzong Wang, Xinhui Liu, Liangxin Gao, et al. Idrid: Diabetic retinopathy–segmentation and grading challenge. *Medical image analysis*, 59:101561, 2020.
26. M Karthick and D Sohier. Aptos 2019 blindness detection. *Kaggle https://kaggle.com/competitions/aptos2019-blindness-detection Go to reference in chapter*, 2019.
27. Yi Zhou, Boyang Wang, Lei Huang, Shanshan Cui, and Ling Shao. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, 2020.
28. Qijie Wei, Xirong Li, Weihong Yu, Xiao Zhang, Yongpeng Zhang, Bojie Hu, Bin Mo, Di Gong, Ning Chen, Dayong Ding, et al. Learn to segment retinal lesions and beyond. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7403–7410. IEEE, 2021.
29. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
30. Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.