# Vision-Language Models Guided Graph Concept Reasoning for Interpretable Diabetic Retinopathy Diagnosis

**Qihao Xu[1,2], Xiaoling Luo[1*], Yuxin Lin[2], Chengliang Liu[3], Yongting Hu [2], Jinkai Li[4], Xinheng Lyu[1,5], Yong Xu[2]**

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen
[3]Department of Computer and Information Science, University of Macau
[4]College of Computer Science and Cyber Security, Chengdu University of Technology
[5]School of Computer Science, University of Nottingham Ningbo China
xqh51199597@outlook.com, xiaolingluoo@outlook.com, linyuxin6688@gmail.com, liucl1996@163.com, huyongting08@163.com, smilelijinkai@gmail.com, Xinheng.LYU@nottingham.edu.cn, laterfall@hit.edu.cn

## Abstract

Deep neural networks (DNNs) have significantly advanced diabetic retinopathy (DR) diagnosis, yet their black-box nature limits clinical acceptance due to a lack of interpretability. Concept bottleneck model (CBM) offers a promising solution by enabling concept-level reasoning and test-time intervention, with recent DR studies modeling lesions as concepts and grades as outcomes. However, current methods often ignore relationships between lesion concepts across different DR grades and struggle when fine-grained lesion concepts are unavailable, limiting their interpretability and real-world applicability. To bridge these gaps, we propose VLM-GCR, a vision-language model guided graph concept reasoning framework for interpretable DR diagnosis. VLM-GCR emulates the diagnostic process of ophthalmologists by constructing a grading-aware lesion concept graph that explicitly models the interactions among lesions and their relationships to disease grades. In concept-free clinical scenarios, our method introduces a vision-language guided dynamic concept pseudo-labeling mechanism to mitigate the challenges of existing concept-based models in fine-grained lesion recognition. Additionally, we introduce a multi-level intervention method that supports error correction, enabling transparent and robust human-AI collaboration. Experiments on two public DR benchmarks show that VLM-GCR achieves strong performance in both lesion and grading tasks, while delivering clear and clinically meaningful reasoning steps.

## Introduction

Diabetic retinopathy (DR) is a common complication of diabetes that damages the blood vessels in the retina, potentially leading to vision loss or blindness. Early detection and timely treatment are essential to managing the progression of this condition and preserving eyesight. Internationally, ophthalmologists classify DR into five stages (grades): normal, mild, moderate, severe, and proliferative DR (PDR). The lesions of DR include hard exudates (EX), soft exudates (SE), hemorrhage (HE), microaneurysms (MA), vitreous hemorrhage (VH), and vitreous opacity (VO), etc.

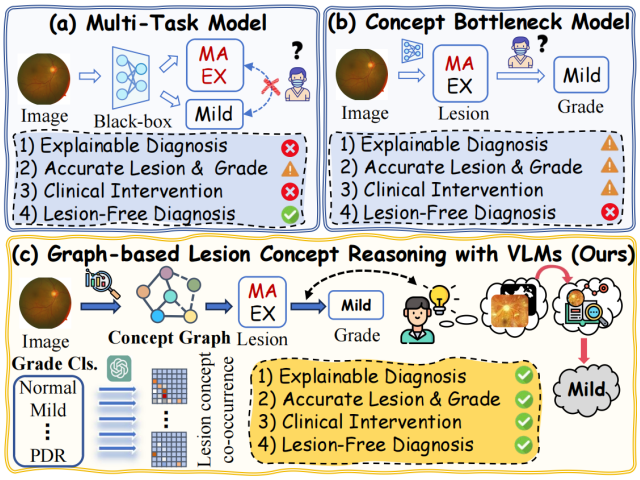



Figure 1: A brief illustration and comparative analysis of DR diagnostic model pipelines. Here, ✓ indicates the method fully satisfies the requirement, △ means the method partially satisfies it, and × means it is not satisfied.

(Wilkinson et al. 2003). Grounded in clinical expertise, the presence, type, and severity of these lesions are key indicators used to determine the DR grade. Based on these ophthalmic diagnostic criteria and clinical requirements, the DR diagnosis task aims to predict the corresponding DR grade from fundus images and their observed retinal lesions.

Deep Neural Networks (DNNs) have achieved remarkable performance in DR diagnosis (Luo et al. 2021, 2023; Lin et al. 2025; Hao, Gao, and Hu 2025), but their opaque decision-making process lacks interpretability. As shown in Fig. 1(a), although the multi-task model can predict both DR lesions and grades, it overlooks the intrinsic relationships between them, leading to unreliability in clinical decisions. Previous studies have employed post-hoc explanations for black-box models by feature visualizations (Selvaraju et al. 2017; Hao, Zhang, and Zhang 2025) to identify potential lesions (He et al. 2020; Zhang et al. 2021). However, empirical findings indicate that the feature maps from these ap-

*Corresponding Author: Xiaoling Luo

proaches are unreliable for incorrectly classified samples.

Unlike post-hoc explanations, medical interpretable models place greater emphasis on making the entire diagnostic process comprehensible, enabling greater trustworthiness for both patients and ophthalmologists. Following this principle, Concept Bottleneck Model (CBM) (Koh et al. 2020) proposes an approach: first providing human-interpretable concepts and then predicting the final class label based on these concepts, as illustrated in Fig. 1(b). One key advantage of this model is that it allows decisions to be influenced by intervening on concepts, which is valuable in clinical medical diagnosis. Consequently, researchers have started applying these methods to medical image analysis (Hu et al. 2024; Yan et al. 2023; Wen et al. 2024; Gao et al. 2024). These methods commonly treat clinical signs (e.g., lesions in the case of DR) as concepts and the diagnostic conclusions (e.g., DR grades) as classification outcomes. There are several CBM variants (Espinosa et al. 2022; Yuksekgonul, Wang, and Zou 2022; Xu et al. 2024b; Liu et al. 2025; Prasse et al. 2025), each addressing interpretability challenges in different domains. These offer valuable insights into deploying concept-based models for interpretable DR diagnosis.

Despite recent advances in concept-based models, interpretable DR diagnosis remains constrained by several key limitations **challenges**. Existing CBM variants treat lesion concepts as conditionally independent given the grade labels, which overlooks their correlations across DR stages. For example, MA and HE frequently appear together in moderate grade but not in mild grade. Moreover, annotating fine-grained DR lesion concepts demands extensive expert labor and time. Under concept-free paradigms, CLIP- and LLM-based CBMs (Lin et al. 2023; Yang et al. 2023) fail to capture these lesion features, undermining both interpretability and diagnostic accuracy. In addition, current intervention methods focus solely on concepts, making it hard to meet the multi-level intervention requirements of clinical knowledge, lesions, and grades in DR diagnosis, thereby hindering the establishment of a robust error correction mechanism (Wang et al. 2021).

To address the above issues, we propose a **V**ision-**L**anguage **M**odels Guided **G**raph **C**oncept **R**easoning framework (**VLM-GCR**) for interpretable DR diagnosis as shown in Fig. 1(c), which simulates the diagnostic thinking of ophthalmologists by constructing a reasoning chain from DR clinical knowledge to lesion concepts and ultimately to grading conclusions. In this work, there are three key components: **1)** we propose a learnable grading-aware lesion concept graph for each fundus image to capture inter-concept relationships under specific grading conditions. This graph enables effective concept interaction and enhances interpretability through structured lesion information. To build it, we first construct a grade-specific concept correlation matrix using either VLMs or co-occurrence statistics. The final graph structure is derived based on auxiliary grades. **2)** During concept-free training, our VLM-GCR introduces a vision-language guided dynamic concept pseudo-labeling method to generate DR lesion concept labels. This approach effectively addresses the limitations of existing methods in capturing fine-grained lesion concepts. **3)** Our reasoning framework adopts a multi-level test-time intervention paradigm that is better suited for clinical human-machine interaction. It enables ophthalmologists or VLMs to intervene at any single level, such as lesion concepts, inter-concept relationships, or grades, and then the model automatically adjusts the other two levels. Extensive experiments on two DR datasets prove that VLM-GCR achieves strong performance in both lesion classfication and grading tasks under various concept supervision settings. Our contributions are summarized as follows:

- Motivated by the diagnostic thinking of ophthalmologists, we propose VLM-GCR for interpretable DR diagnosis which mimics clinical reasoning through concept-level modeling. The proposed grading-aware lesion concept graph dynamically models the correlations between concepts under different grades, enabling concept interaction and enhancing interpretability.
- We propose a vision-language guided dynamic concept pseudo-labeling mechanism for concept-free training, addressing the dilemma faced by existing models in fine-grained lesion concept recognition.
- We further introduce a novel multi-level intervention mechanism that enables targeted error correction at lesion, grade, or relationship levels, facilitating transparent and robust human-AI collaboration.

# Preliminary

## Problem Definition

Given a supervised DR diagnosis dataset $\mathcal{D} = \left\{\mathbf{X}^{(j)}, \mathbf{C}^{(j)}, \mathbf{Y}^{(j)}\right\}_{j=1}^{S}$ with $S$ samples, $N$ lesion concepts, and $K$ grading categories, where the $j$-th sample consists of the fundus image $\mathbf{X}^{(j)} \in \mathcal{X}$, lesion concepts label $\mathbf{C}^{(j)} \in \mathcal{C} \subset \{0,1\}^N$, and the DR grading label $\mathbf{Y}^{(j)} \in \mathcal{Y} \subset \{0,1\}^K$. It is noted that $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{C}$ are the spaces of input, $K$-dimensional one-hot vectors and $N$-dimensional vectors. We denote $\mathbf{C}_n \in \mathcal{C}$ and $\mathbf{Y}_k \in \mathcal{Y}$ as the vectors of the $n$-th concept and the $k$-th grade, respectively. The interpretable DR diagnosis task aims to train a network that accurately predicts lesion concepts $\mathbf{C}$ and DR grade $\mathbf{Y}$ under both concept-supervised and concept-free settings, while ensuring clinical reliability and integrated interpretability for real-world medical applications. Here, concept-free indicates that no lesion concept labels are provided during training, whereas concept-supervised refers to settings where such annotations are available.

## Concept-Based Models

As a representative interpretable model, CBM follows a two-step process: it first maps the input space $\mathcal{X}$ into the concept space $\mathcal{C}$, and then transforms the concept representations into the class space $\mathcal{Y}$, i.e., $\mathcal{X} \rightarrow \mathcal{C} \rightarrow \mathcal{Y}$. However, due to the excessive information compression at the concept bottleneck, CBM suffers from inferior accuracy compared to black-box models. To balance interpretability and accuracy, Concept Embedding Model (CEM) (Espinosa et al. 2022) and the Post-hoc Concept Bottleneck Model (PCBM-h) (Yuksekgonul 2022) propose positive-negative concept

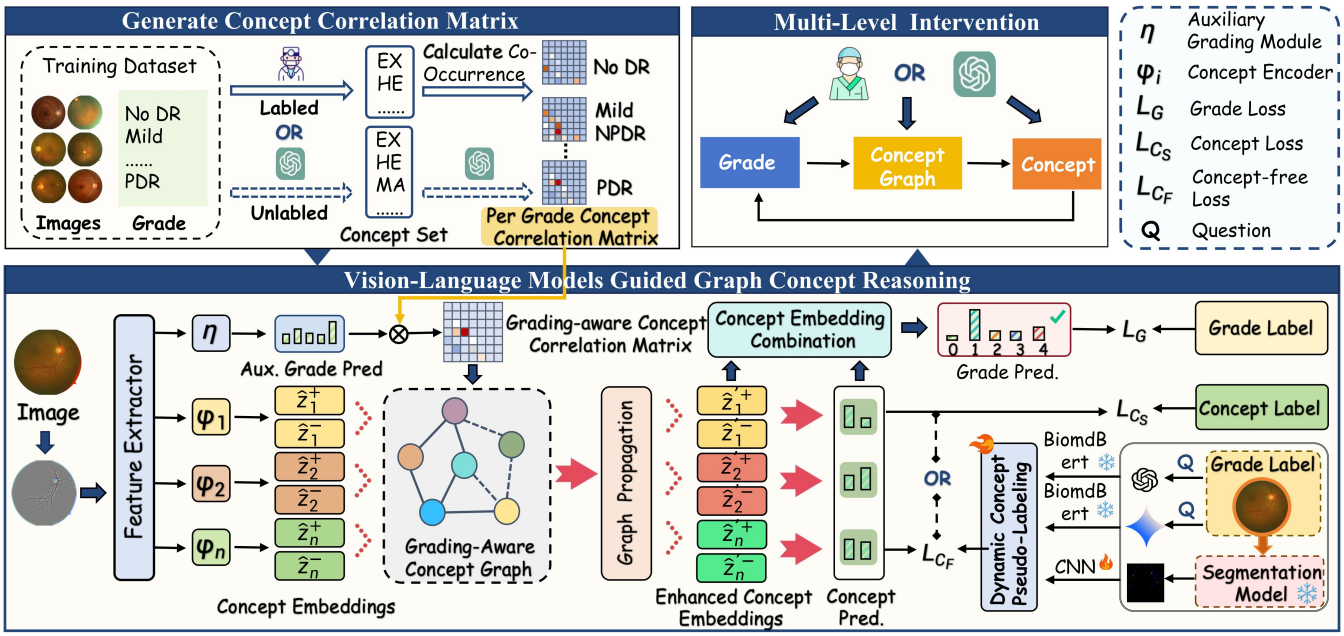



Figure 2: Our VLM-GCR framework consists of grade-specific concept correlation modeling, graph-based concept reasoning, and multi-level intervention. The use of VLMs contributes to the pipeline of models during training and inference.

semantics and residual adaptation methods. Kim et al. (Kim et al. 2023) and Xu et al. (Xu et al. 2024b) propose PCBM and ECBM, introducing probabilistic and energy-based theories for concept learning. Although CBM and its variants have been applied to interpretable medical analysis (Yan et al. 2023; Wen et al. 2024; Gao et al. 2024), they face the challenges presented in Sec. 1 when used in DR diagnosis.

## Method

In this section, our framework follows a three-stage reasoning pipeline, including grade-specific concept correlation modeling, graph-based interpretable reasoning, and multi-level intervention at test time. This interpretable reasoning chain simulates the diagnostic thinking of ophthalmologists from DR clinical knowledge to lesion concepts and ultimately to grading conclusions. In addition, the VLMs are involved in all stages. During concept-free training, we propose a dynamic concept pseudo-labeling mechanism using VLMs to address the difficulty of capturing lesion concepts.

### Grade-specific Concept Correlation Modeling

To bridge the relationships between lesion concepts for each grade, we propose a grade-specific concept correlation modeling approach as shown in Fig. 2. In this work, we generate DR lesion concept co-occurrence matrices for each grade to represent their correlations, with dedicated computation methods designed for both concept-supervised and concept-free settings. These matrices are then used to construct the grading-aware concept graph, where concepts serve as nodes and their pairwise co-occurrences as edges. The design of this graph is aligned with DR clinical guidelines and ensures the correctness and interpretability of concept modeling.

In the concept-supervised setting, we obtain the co-occurrence matrix by calculating the frequency with which any two lesion concepts co-occur in each grade in the training dataset. Specifically, we define $\mathcal{P}(i, j|k)$ as the probability that the $i$-th and $j$-th concepts co-occur under grade $k$. To reduce computational cost, we construct the grade $k$ concept correlation matrix $\mathbf{M}^{(k)} \in \mathbb{R}^{N\times N}$ by:

$$\mathbf{M}_{ij}^{(k)} = \mathcal{P}(i, j|k) = \frac{\sum_{u=1}^{S} \mathbf{Y}_k^{(u)} \mathbf{C}_i^{(u)} \mathbf{C}_j^{(u)}}{\sum_{u=1}^{S} \mathbf{Y}_k^{(u)}} = \frac{(\mathbf{Y}^\top \mathbf{C})^\top (\mathbf{Y}^\top \mathbf{C})}{\mathbf{Y}^\top \mathbf{1}_N}. \quad (1)$$

Where $k \in \{0, 1, \ldots, K-1\}$, and $\mathbf{1}_N \in \mathbb{R}^{N\times 1}$ denotes an all-ones vector of length $N$. The diagonal elements of $\mathbf{M}^{(k)}$ are set to zero. Moreover, due to its symmetry, the $\mathbf{M}^{(k)}$ is simplified to a lower triangular form. In the concept-free setting, inspired by recent work (Hu et al. 2024), we first employ GPT-4.1 to generate DR-related lesion concepts. Leveraging GPT-4.1 extensive knowledge of DR diagnosis, we then prompt it to estimate the approximate co-occurrence probability $\mathbf{M}_{ij}^{(k)}$ between lesion concepts $i$ and $j$ under DR grade $k$. The resulting matrix $\mathbf{M}$ is used to initialize the concept graph. Details of the prompting process are presented in the appendix.

### Graph-based Concept Reasoning

**Graph Concept Representation Learning** To build an interpretable DR diagnostic chain from symptom analysis to grading conclusions, we propose a graph-based lesion concept reasoning framework as shown in Fig. 2. Given fundus images $\hat{\mathbf{X}}$ as the input, we first apply preprocessing methods, including automatic cropping and contrast enhancement, to

obtain the processed input $\hat{\mathbf{X}}'$. Subsequently, feeding $\hat{\mathbf{X}}'$ into the backbone network to extract hidden features $\hat{\mathbf{H}}$. Then, $\hat{\mathbf{H}}$ flows into two branches: one branch produces the lesion concept embeddings, while the other generates the grading-aware concept correlation matrix $\mathbf{M}^{\hat{\mathbf{x}}}$.

In the first branch, the concept encoders $\{\varphi_i\}_{i=1}^N$, consisting of fully connected layers, transform $\hat{\mathbf{H}}$ into concept embeddings $\hat{\mathbf{Z}} = \{\hat{\mathbf{Z}}_i\}_{i=1}^N$ and split them into positive and negative parts $\{\hat{\mathbf{Z}}_i^+\}_{i=1}^N, \{\hat{\mathbf{Z}}_i^-\}_{i=1}^N$ along the channel, representing embeddings that support and oppose the presence of concept. Here, $\hat{\mathbf{Z}}_i$ denotes the $i$-th concept embedding. We formulate this process as:

$$\{\hat{\mathbf{Z}}_i\}_{i=1}^N = \{[\hat{\mathbf{Z}}_i^+, \hat{\mathbf{Z}}_i^-]\}_{i=1}^N = \{\text{Split}(\varphi_i(\hat{\mathbf{H}}))\}_{i=1}^N. \quad (2)$$

In the other branch, we compute the auxiliary DR grade $\hat{\mathbf{A}} = \eta(\hat{\mathbf{H}})$ by an auxiliary classifier $\eta$. Using the maintained concept correlation matrix $\mathbf{M} = [\mathbf{M}^{(0)}, \mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(K-1)}]$ and $\hat{\mathbf{A}}$, the grading-aware concept correlation matrix $\mathbf{M}^{\hat{\mathbf{x}}}$ is calculated in the form of weighted summation:

$$\mathbf{M}^{\hat{\mathbf{x}}} = \sum_{k=0}^{K-1} \hat{\mathbf{A}}_k \mathbf{M}^{(k)}, \quad \hat{\mathbf{A}}_k = \frac{\exp(\hat{\mathbf{A}}_k)}{\sum_{j=0}^{K-1} \exp(\hat{\mathbf{A}}_j)}. \quad (3)$$

Next, we construct a learnable grading-aware concept graph $\mathcal{G} = \{\mathbf{V}, \mathbf{E}\}$, where the nodes set $\mathbf{V}$ contains $N$ concept embeddings $\{\hat{\mathbf{Z}}_i\}_{i=1}^n$ and the edge set $\mathbf{E}$ represents relationships between adjacent concepts. Naturally, the edge weights are initialized by $\mathbf{M}^{(\hat{\mathbf{x}}')}$, enabling the graph to dynamically model inter-concept relationships conditioned on the auxiliary grade.

Then, we adopt a graph message passing mechanism based on Gated Recurrent Units (GRU) (Li et al. 2015; Chen et al. 2019) to enable interactions among lesion concept features. Given the $\{\hat{\mathbf{Z}}_i\}_{i=1}^N$, the initial node states are computed as: $\mathbf{S}^{(0)} = l_{\text{init}} \cdot \{\hat{\mathbf{Z}}_i\}_{i=1}^n$, where $l_{\text{init}}$ is a learnable weight matrix. At each step $t$, the nodes aggregate messages $\mathbf{m}^{(t)}$ from neighbors via $\mathbf{M}^{(\hat{\mathbf{x}}')}$:

$$\mathbf{m}^{(t)} = \text{ReLU}\left(\mathbf{M}^{(\hat{\mathbf{x}}')}\mathbf{S}^{(t-1)}\right). \quad (4)$$

The nodes are updated through GRU for graph propagation:

$$\mathbf{S}^{(t)} = \text{GRU}(\mathbf{S}^{(t-1)}, \mathbf{m}^{(t)}; \Theta), \quad (5)$$

where GRU enables selective integration of relational concept features, and $\Theta$ denotes the learnable parameters of it. After $T$ iterations ($T = 3$ in our model), we obtain the enhanced concept embeddings $\{\hat{\mathbf{Z}}_i'\}_{i=1}^n = \mathbf{S}^{(T)}$. In this way, lesion concepts with higher co-occurrence probabilities under the current grade engage in more interactions, thereby optimizing their visual representations.

**Grading through Concept Reasoning** To balance interpretability and diagnostic accuracy, we introduce a reasoning approach based on the combination of concept embeddings for interpretable DR grading. The enhanced concept embeddings $\{\hat{\mathbf{Z}}_i'\}_{i=1}^n$ are firstly fed into the concept classifier to obtain the predictions for each lesion concept. Based on the predicted concept labels $\{\hat{\mathbf{C}}_i\}_{i=1}^n$, along with the corresponding positive and negative concept embeddings $\{\hat{\mathbf{Z}}_i'^+\}_{i=1}^n$ and $\{\hat{\mathbf{Z}}_i^-\}_{i=1}^n$, the final combined concept embedding $\overline{\mathbf{Z}}$ is computed as:

$$\overline{\mathbf{Z}} = \text{Concat}[\{\alpha_i \cdot \hat{\mathbf{Z}}_i^+ + (1-\alpha_i) \cdot \hat{\mathbf{Z}}_i^-\}_{i=1}^N], \quad (6)$$

where Concat$[\cdot,\cdot]$ denotes channel-wise concatenation, and the concept prediction probability is $\alpha_i = \frac{1}{1+\exp(-\hat{\mathbf{C}}_i)}$. Finally, the model uses $\overline{\mathbf{Z}}$ to predict the DR grade $\hat{\mathbf{Y}}$ through the grading classifier. As a result, our VLM-GCR establishes an interpretable reasoning chain from the diagnosis-aligned concept graph, through concept analysis, to final grading.

## Vision-Language Models Guided Training

**Dynamic Concept Pseudo-Labeling** Due to the scarcity of annotated lesion concepts in DR, it is essential to design training methods that do not rely on concept supervision for clinical diagnosis. Under the concept-free setting, prior approaches employ contrastive learning based on CLIP to train concept-based models. However, this approach fails to capture subtle lesion concepts. In our preliminary exploration, we investigate the application of advanced VLMs (e.g., GPT-4.1) for automatic DR lesion annotation, but find that this method is prone to hallucinations. To address the issues caused by single VLM-based annotation, we propose a dynamic concept pseudo-labeling strategy inspired by the mixture-of-experts (MoE) model (Riquelme et al. 2021).

For each fundus image in the training set, we construct a conditional prompt based on its corresponding grading label. This prompt, along with the image, is then fed into two large vision-language models (GPT-4.1 and Gemini 2.5 Flash) to generate diagnostic-related textual outputs, denoted as $\{\mathbf{R}_i\}_{i=1}^2$. Meanwhile, pseudo lesion segmentation masks $\mathbf{R}_3$ are obtained by applying the pretrained DR segmentation model (Xu et al. 2024a). Subsequently, $\{\mathbf{R}_i\}_{i=1}^2$ and $\mathbf{R}_3$ are processed by two frozen text encoders $\mathcal{F}_1, \mathcal{F}_2$ (Gu et al. 2021) and one non-frozen image encoder $\mathcal{F}_3$, respectively, to generate their corresponding pseudo concept features $\mathcal{F}_i(\mathbf{R}_i) \in \mathbb{R}^D$. We build trainable shared-expert classifiers $\{\sigma_i\}_{i=1}^3$ for the three features to generate their concept pseudo labels. These labels are then fused using a dynamic approach, followed by threshold filtering. This labeling process is denoted as:

$$\mathbf{P}_j^{(i)} = \begin{cases} 1, & \left[\sum_{k=1}^{3}\sum_{i=1}^{3} \mathbf{L}^k \cdot \sigma_i[\mathcal{F}_k.(\mathbf{R}_k)]\right]_j^{(i)} \geq \theta, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where $\mathbf{P}$ is the binary concept pseudo label matrix, $\theta$ ($\theta = 0.4$) is the confidence threshold, and $\mathbf{L}^k$ represents the adaptive weight of the $k$-th expert, computed via a Laplace gating router network (Han et al. 2024):

$$\mathbf{L}^k = \frac{\exp(h_k(\mathcal{F}_k(\mathbf{R}_k)))}{\sum_{l=1}^{3}\exp(h_l(\mathcal{F}_l(\mathbf{R}_l)))}, h_k = -\|W_k - \mathcal{F}_k(\mathbf{R}_k))\|_2. \quad (8)$$

Here, $W_k \in \mathbb{R}^D$ is the learnable parameter. The Laplace gating function measures the dissimilarity between each expert and the input using the Euclidean term $\exp(-\|W_k -$

| Task | Data | Metric | Multi-Task (CVPR 19) | CBM (ICML 20) | CEM (NIPS 22) | PCBM (ICML 23) | LEN (AI 23) | ECEM (ICLR 24) | Evi-CEM (MICA 24) | CLAT (TMI 24) | **VLM-GCR (Ours)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Concept Supervision | DDR | ACC | $81.28_{\pm 0.82}$ | $80.78_{\pm 1.44}$ | $81.57_{\pm 1.59}$ | $81.01_{\pm 2.12}$ | $81.14_{\pm 1.06}$ | $82.10_{\pm 0.42}$ | $82.23_{\pm 1.26}$ | $82.32_{\pm 1.14}$ | $\mathbf{83.10}_{\pm 0.74}$ |
| | | AUC | $93.52_{\pm 0.32}$ | $93.14_{\pm 0.53}$ | $93.22_{\pm 0.69}$ | $93.15_{\pm 1.11}$ | $93.38_{\pm 0.68}$ | $94.01_{\pm 0.79}$ | $93.56_{\pm 0.89}$ | $94.09_{\pm 0.67}$ | $\mathbf{94.74}_{\pm 0.51}$ |
| | | AUPR | $89.25_{\pm 0.35}$ | $87.35_{\pm 0.55}$ | $88.42_{\pm 0.57}$ | $87.99_{\pm 0.97}$ | $89.01_{\pm 0.78}$ | $88.79_{\pm 0.55}$ | $88.39_{\pm 0.71}$ | $88.51_{\pm 0.55}$ | $\mathbf{89.82}_{\pm 0.73}$ |
| | | Kappa | $73.52_{\pm 0.35}$ | $74.17_{\pm 1.89}$ | $\mathbf{75.21}_{\pm 2.13}$ | $73.35_{\pm 1.34}$ | $74.14_{\pm 1.41}$ | $74.19_{\pm 0.76}$ | $74.11_{\pm 1.41}$ | $74.52_{\pm 2.04}$ | $75.15_{\pm 1.15}$ |
| | | F1 | $71.44_{\pm 0.20}$ | $67.78_{\pm 1.96}$ | $71.73_{\pm 1.51}$ | $67.92_{\pm 1.51}$ | $71.50_{\pm 0.43}$ | $72.01_{\pm 1.01}$ | $71.88_{\pm 1.34}$ | $72.17_{\pm 1.11}$ | $\mathbf{73.58}_{\pm 0.86}$ |
| | MFIDDR | ACC | $73.74_{\pm 0.37}$ | $73.02_{\pm 0.40}$ | $74.11_{\pm 0.45}$ | $73.75_{\pm 1.21}$ | $74.02_{\pm 0.39}$ | $74.55_{\pm 0.78}$ | $74.21_{\pm 1.12}$ | $74.32_{\pm 0.67}$ | $\mathbf{75.47}_{\pm 0.61}$ |
| | | AUC | $87.30_{\pm 0.99}$ | $86.22_{\pm 0.12}$ | $86.71_{\pm 0.79}$ | $86.01_{\pm 0.71}$ | $86.78_{\pm 0.43}$ | $87.52_{\pm 0.52}$ | $87.31_{\pm 0.17}$ | $87.81_{\pm 0.37}$ | $\mathbf{88.91}_{\pm 0.67}$ |
| | | AUPR | $76.31_{\pm 0.82}$ | $76.51_{\pm 0.82}$ | $77.02_{\pm 0.85}$ | $75.23_{\pm 0.62}$ | $76.50_{\pm 0.43}$ | $77.03_{\pm 1.01}$ | $76.89_{\pm 0.68}$ | $76.73_{\pm 0.71}$ | $\mathbf{78.19}_{\pm 0.90}$ |
| | | Kappa | $51.27_{\pm 0.60}$ | $51.15_{\pm 0.43}$ | $52.04_{\pm 0.75}$ | $50.97_{\pm 1.10}$ | $51.71_{\pm 0.40}$ | $52.04_{\pm 0.21}$ | $51.44_{\pm 0.98}$ | $51.97_{\pm 0.96}$ | $\mathbf{52.51}_{\pm 0.71}$ |
| | | F1 | $55.23_{\pm 2.10}$ | $55.94_{\pm 0.62}$ | $56.37_{\pm 1.27}$ | $55.12_{\pm 1.12}$ | $55.53_{\pm 1.10}$ | $56.03_{\pm 1.04}$ | $56.02_{\pm 1.20}$ | $56.11_{\pm 0.62}$ | $\mathbf{56.78}_{\pm 1.51}$ |

Table 1: Performance comparison of DR grading with concept supervision across three random seeds. (mean ± std, Unit: %)

| Task | Data | Metric | PCBM-h (ICLR 23) | LFCBM (ICLR 23) | Med-MICN (NIPS 24) | SSCBM (ICCVW 25) | **VLM-GCR (Ours)** |
|---|---|---|---|---|---|---|---|
| Concept Free | DDR | ACC | $80.42_{\pm 1.59}$ | $79.93_{\pm 2.00}$ | $80.79_{\pm 1.02}$ | $80.54_{\pm 0.79}$ | $\mathbf{81.82}_{\pm 0.58}$ |
| | | AUC | $91.73_{\pm 2.82}$ | $92.15_{\pm 1.42}$ | $91.98_{\pm 2.11}$ | $92.31_{\pm 1.71}$ | $\mathbf{93.17}_{\pm 0.81}$ |
| | | AUPR | $85.22_{\pm 1.74}$ | $85.20_{\pm 1.52}$ | $86.31_{\pm 1.34}$ | $86.02_{\pm 1.14}$ | $\mathbf{87.68}_{\pm 0.65}$ |
| | | Kappa | $70.93_{\pm 2.48}$ | $71.52_{\pm 3.51}$ | $72.30_{\pm 0.77}$ | $71.30_{\pm 2.04}$ | $\mathbf{72.75}_{\pm 0.93}$ |
| | | F1 | $69.59_{\pm 2.10}$ | $69.87_{\pm 2.41}$ | $69.87_{\pm 2.11}$ | $69.02_{\pm 1.79}$ | $\mathbf{71.64}_{\pm 1.73}$ |
| | MFIDDR | ACC | $72.95_{\pm 0.64}$ | $72.46_{\pm 0.72}$ | $73.20_{\pm 0.19}$ | $71.98_{\pm 0.51}$ | $\mathbf{74.25}_{\pm 0.30}$ |
| | | AUC | $85.91_{\pm 0.77}$ | $85.71_{\pm 1.11}$ | $85.74_{\pm 0.49}$ | $85.34_{\pm 0.90}$ | $\mathbf{87.55}_{\pm 0.65}$ |
| | | AUPR | $75.01_{\pm 1.03}$ | $74.73_{\pm 1.32}$ | $75.24_{\pm 0.47}$ | $74.61_{\pm 1.21}$ | $\mathbf{77.10}_{\pm 0.58}$ |
| | | Kappa | $50.14_{\pm 0.76}$ | $50.09_{\pm 1.61}$ | $50.77_{\pm 0.65}$ | $49.91_{\pm 1.09}$ | $\mathbf{51.75}_{\pm 0.87}$ |
| | | F1 | $54.32_{\pm 1.23}$ | $54.11_{\pm 2.17}$ | $54.72_{\pm 1.09}$ | $53.87_{\pm 1.89}$ | $\mathbf{55.85}_{\pm 1.19}$ |

Table 2: Performance comparison of DR grading in the concept-free scenario across three random seeds. (mean ± std, Unit: %)

$\mathcal{F}_k(\mathbf{R}_k))\|_2)$. Due to its bounded nature, this term helps prevent extreme weight distributions. This effective approach significantly improves concept recognition under the concept-free setting, effectively addressing the limitations of existing methods in capturing fine-grained lesion concepts.

**Loss function** The loss function of VLM-GCR $\mathcal{L}$ combines a final grading loss $\mathcal{L}_{\mathrm{g}}$, an auxiliary grading loss $\mathcal{L}_{\mathrm{a}}$, and a lesion concept loss $\mathcal{L}_{\mathrm{c}}$, and is formulated as:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\mathrm{g}} + \alpha\mathcal{L}_{\mathrm{a}} + \beta\mathcal{L}_{\mathrm{c}} \\ &= \mathrm{CE}(\hat{\mathbf{Y}}, \mathbf{Y}) + \alpha \cdot \mathrm{CE}(\hat{\mathbf{A}}, \mathbf{Y}) \\ &\quad + \beta \cdot \begin{cases} \mathrm{BCE}(\hat{\mathbf{C}}, \mathbf{P}), & \text{(concept-free)} \\ \mathrm{BCE}(\hat{\mathbf{C}}, \mathbf{C}), & \text{(concept-supervised)} \end{cases}\end{aligned} \tag{9}$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes the cross-entropy loss used for DR grading, and $\mathrm{BCE}(\cdot,\cdot)$ is the binary cross-entropy loss applied to multi-label lesion concept prediction. $\alpha$ and $\beta$ are hyperparameters that balance the $\mathcal{L}_{\mathrm{a}}$ and $\mathcal{L}_{\mathrm{c}}$, respectively.

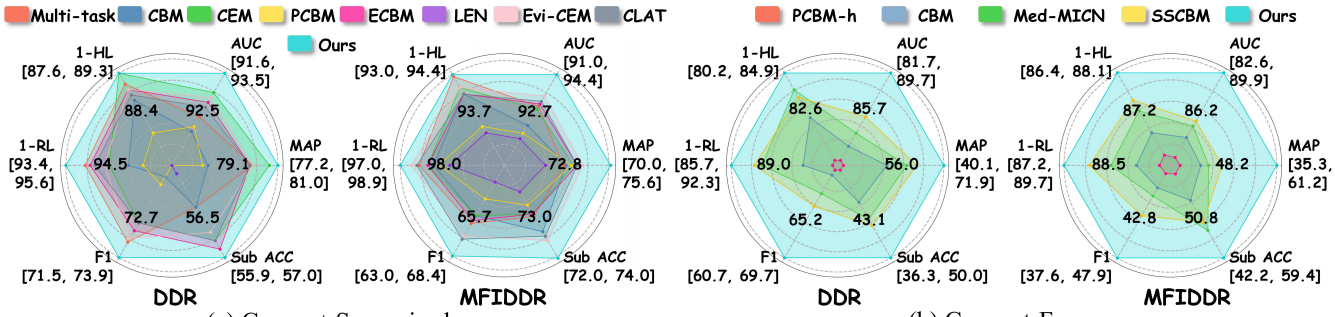
## Multi-Level Intervention

Like most concept-based models, VLM-GCR supports clinical test-time intervention, as illustrated in Fig. 2. The key difference is that the former is limited to intervening lesion concepts (symptoms) $\mathcal{C}$ to correct DR grade $\mathcal{Y}$, i.e., $\mathcal{C} \rightarrow \mathcal{Y}$. In contrast, VLM-GCR also allows ophthalmologists or VLMs to intervene in the diagnostic reasoning process in two ways. First, by modifying the diagnostic grade, they can adjust the edges of the grading-aware concept graph to correct recognition errors in lesion concepts, following the path $\mathcal{Y} \rightarrow \mathcal{G} \rightarrow \mathcal{C}$. Alternatively, they can directly edit the concept graph edges based on domain-specific DR knowledge, thereby enhancing both lesion concept classification and grading accuracy, i.e., $\mathcal{G} \rightarrow \mathcal{C} \rightarrow \mathcal{Y}$. Specifically, during the inference, we can input the known grade (one-hot encoding), concepts, or the grading-aware concept correlation matrix $\mathbf{M}^{(\mathbf{x}')}$ into the model to modify the original grade or concepts. According to Eq. 4 and 5, the concept correlation error is amplified in the message passing of GNN, leading to deviations in node representations from the actual topology. It proves the impacts of directly or indirectly adjusting concept relations for correcting grading predictions.

# Experiments

## Experimental Setups

**Datasets** We conduct experiments on two DR datasets, DDR (Li et al. 2019) and MFIDDR (Luo et al. 2023). The full DDR dataset comprises 13,673 fundus images with grade labels (757 of which have segmentation labels). Following (Wen et al. 2024), we select 2,334 of them (1,045 with grade 0) for experiments, which have both grade and lesion concept labels. This dataset is randomly partitioned into 70%, 10%, and 20% for training, validation, and testing, respectively. MFIDDR contains 34,452 single-view fundus images with grade labels (25,848 for training and 8,604 for




(a) Concept-Supervised (b) Concept-Free

Figure 3: Performance comparison of average DR concept classification metrics, consistent with Table 1 settings. (Unit: %)

testing). We invite ophthalmologists to annotate the lesion concept labels for each of the 25,848 images individually, and split these into training and validation sets in an 8:2 ratio. The concepts include EX, HE, MA, SE, VH, and VO.

**Evaluation Metrics** From a performance perspective, the DR diagnosis task aims to achieve both accurate grading and lesion prediction under both concept-free and concept-supervised settings. We use ACC, AUC, AUPR, Macro F1, and Kappa as grading metrics, and Ranking loss (RL), Hamming loss (HL), AUC, mAP, Macro F1, and subset accuracy (Sub ACC) as lesion concept classification metrics. Notably, no interventions are applied in the comparison experiments.

**Baselines** In the concept-supervised setting of the DR diagnosis task, we compare our VLM-GCR with a non-interpretable Multi-Task model (Liu, Johns, and Davison 2019) and several concept-based interpretable models, including CBM (Koh et al. 2020), CEM (Espinosa et al. 2022), PCBM (Kim et al. 2023), LEN (Ciravegna et al. 2023), ECBM (Xu et al. 2024b), Evi-CEM (Gao et al. 2024), and CLAT (Wen et al. 2024). In the concept-free setting, the compared methods contain PCBM-h (Yuksekgonul, Wang, and Zou 2022), LFCBM (Oikarinen et al. 2023), Med-MICN (Hu et al. 2024), and SSCBM (Hu et al. 2025).

**Implementation Details** All experiments are conducted on an NVIDIA RTX 4090 GPU. Fundus images are input at 224×224, and training lasts 100 epochs in 8 batch sizes. The Adam optimizer is used with an initial learning rate of 0.00005, which dynamically adjusts through a cosine annealing scheduler. To ensure fair comparisons, all methods use VGG16 (Simonyan and Zisserman 2015) as the backbone and are evaluated across three different random seeds. Following (Espinosa et al. 2022), we test both DR grading and lesion concept classification by using the model weights that yield the best grading accuracy on the validation set.

## Experimental Results and Analysis

**Comparison with Advanced Methods** In this subsection, we compare our method with twelve other approaches for DR diagnosis across two datasets. Of these, eight are used for concept-supervised evaluation, while the other four are not. According to experimental results for joint DR grading and lesion concept classification in Table 1 and Fig. 3, we can have the following observations. Under concept supervision, VLM-GCR achieves state-of-the-art performance in both DR grading and lesion concept classification. On the DDR and MFIDDR datasets, it improves AUC by 1.23% and 2.07%, and F1 by 2.47% and 0.99% for grading, as well as mAP by 2.00% and 2.80% for lesion classification, respectively. In the concept-free scene, VLM-GCR yields significant improvements in lesion concept classification, with mAP gains of 15.9% and 13.0%, and Sub ACC increases of 6.9% and 9.4%, compared to the average performance of previous methods on the DDR and MFIDDR datasets. These results demonstrate its ability to overcome the limitations of LLM- and CLIP-based interpretable methods in capturing fine-grained lesion concepts under the concept-free setting. Moreover, VLM-GCR supports training in both concept-supervised and concept-free settings, highlighting its generalization capability across different supervision paradigms.

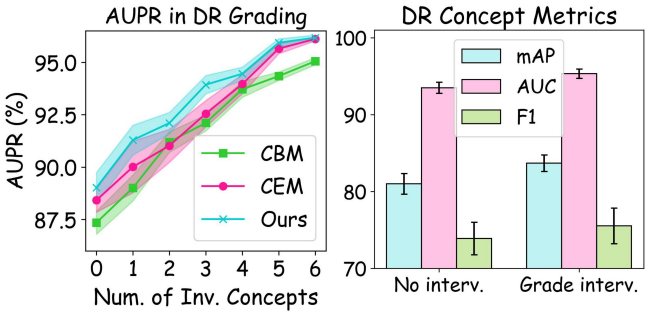



Figure 4: Evaluation of and grading and concept interventions on the DDR dataset over three random seeds.

**Interpretability Analysis** Our model offers three key interpretability advantages over CBM, its variants, and other interpretable methods for DR diagnosis. VLM-GCR enhances interpretability by incorporating a concept graph that provides structured information for concept reasoning. This structure enables the model to simulate the diagnostic process of ophthalmologists, including knowledge grounding, lesion analysis, and decision-making. It also generates fine-grained lesion localization and textual diagnostic reports, as shown in Fig. 5(a), which are valuable for supporting clinical analysis. Furthermore, the lesion concepts, auxiliary grade, and graph structure act as explicitly controllable neurons, al-

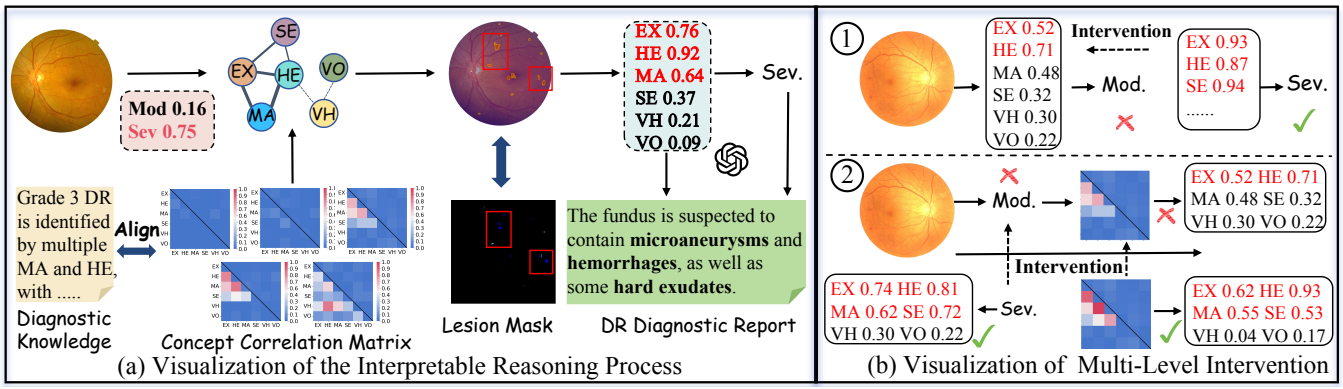



Figure 5: Visualization of the reasoning process and multi-level interventions.

| Concept-supervised | | |
|---|---|---|
| **Method** | **Grade** | **Concept** |
| **VLM-GCR (Ours)** | **94.74** | **93.51** |
| w/o CG | 93.57 | 92.17 |
| w/o GA | 94.11 | 92.49 |
| **Concept-free** | | |
| **VLM-GCR (Ours)** | **93.17** | **89.71** |
| w/o Gemini | 91.97 | 86.19 |
| w/o GPT-4.1 | 91.09 | 84.89 |
| w/o SM | 92.56 | 87.02 |
| w/o DCPL | 92.01 | 82.12 |

Table 3: Ablation results in concept-supervised and concept-free settings on the DDR dataset. The evaluation metric is the average AUC over three random seeds. (Unit: %)

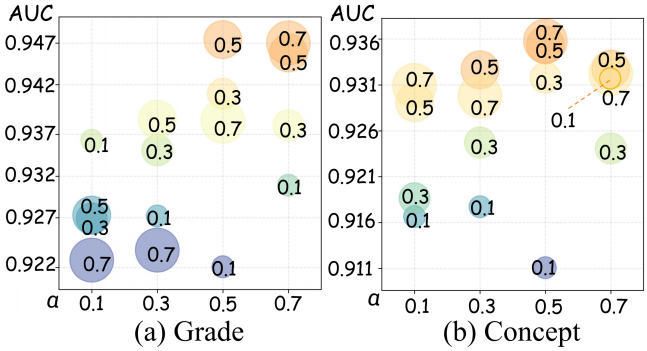



Figure 6: Evaluation of $\alpha$ and $\beta$ on the DDR dataset, with the number inside each circle being the corresponding $\beta$ value.

lowing ophthalmologists to perform targeted interventions at test time. This design promotes transparent and robust human-AI collaboration, as illustrated in Fig. 5(b).

**Evaluation of Multi-Level Intervention** Our VLM-GCR enables interventions at the grading, concept graph, and lesion levels to correct reasoning errors at the other two corresponding levels, as shown in Fig. 5(b). Following (Espinosa et al. 2022), we conduct experiments under different numbers of concept interventions on DDR. In Fig. 4, it can be seen that the DR grading performance improves as the number of concept interventions increases, and our method outperforms both CBM and CEM. Under the grading intervention, our method improves the metrics for lesion classification. which proves the effectiveness of our approach.

## Ablation and Hyperparameter Analysis

We evaluate the performance of our model in two conditions: without concept graph modeling (w/o CG) and with the grade-independent concept graph (w/o GA) under the concept-supervised setting. In Table 3, we observe that the removal of these two parts leads to varying degrees of performance degradation in both the grading and concept AUC metrics. Our dynamic concept pseudo-labeling method integrates VLMs and DR lesion segmentation models. Table 3 validates the effectiveness of the individual models and the dynamic fusion approach. Specifically, compared to directly aggregating the outputs of different models (w/o DCPL), our pseudo-labeling method achieves a 7.59% improvement in lesion concept AUC. In addition, the absence of either VLMs (w/o GPT-4.1; w/o Gemini) or the segmentation model (w/o SM) leads to a performance drop between 2.69% and 4.63% in lesion concept AUC. We evaluate the loss function parameters $\alpha$ and $\beta$. The model achieves the best DR grading and second-best lesion classification AUC when $\alpha = 0.5, \beta = 0.5$, as shown in Fig. 6.

## Conclusion

In this work, we present VLM-GCR for interpretable DR diagnosis that integrates clinical reasoning into model design. By leveraging a grading-aware lesion concept graph, the method enables dynamic concept interaction and improves interpretability. To overcome the challenge of lesion recognition under limited supervision, we introduce a vision-language guided dynamic concept pseudo-labeling mechanism, making concept-free training feasible and effective. Furthermore, the proposed multi-level intervention strategy enhances transparency and robustness by allowing targeted correction at the lesion, grade, and relational levels. Extensive experiments on two datasets demonstrate the effectiveness and generalization capability of our method.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62502320, the Natural Science Foundation of Guangdong Province under Grant No. 2025A1515010184, the project of Shenzhen Science and Technology Innovation Committee under Grant No. JCYJ20240813141424032, the Guangdong Major Project of Basic and Applied Basic Research under Grant 2023B0303000010, and the Scientific Foundation for Youth Scholars of Shenzhen University Grant No. 827-0001083.

## References

Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019. Learning semantic-specific graph representation for multi-label image recognition. In *ICCV*, 522–531.

Ciravegna, G.; Barbiero, P.; Giannini, F.; Gori, M.; Lió, P.; Maggini, M.; and Melacci, S. 2023. Logic explained networks. *Artificial Intelligence*, 314: 103822.

Espinosa, M.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; et al. 2022. Concept embedding models: Beyond the accuracy-explainability trade-off. In *NeurIPS*, volume 35, 21400–21413.

Gao, Y.; Gao, Z.; Gao, X.; Liu, Y.; Wang, B.; and Zhuang, X. 2024. Evidential Concept Embedding Models: Towards Reliable Concept Explanations for Skin Disease Diagnosis. In *MICCAI*, 308–317. Springer.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1): 23.

Han, X.; Nguyen, H.; Harris, C.; Ho, N.; and Saria, S. 2024. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. In *NeurIPS*, volume 37, 67850–67900.

Hao, P.; Gao, W.; and Hu, L. 2025. Embedded feature fusion for multi-view multi-label feature selection. *Pattern Recognition*, 157: 110888.

Hao, P.; Zhang, H.; and Zhang, Y. 2025. Tensor-based Opposing yet Complementary Learning for Multi-view Multi-label Feature Selection. In *ACMMM*, 1822–1831.

He, A.; Li, T.; Li, N.; Wang, K.; and Fu, H. 2020. CABNet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1): 143–153.

Hu, L.; Huang, T.; Xie, H.; Gong, X.; Ren, C.; Hu, Z.; Yu, L.; Ma, P.; and Wang, D. 2025. Semi-supervised concept bottleneck models. In *ICCV*, 2110–2119.

Hu, L.; Lai, S.; Chen, W.; Xiao, H.; Lin, H.; Yu, L.; Zhang, J.; and Wang, D. 2024. Towards Multi-dimensional Explanation Alignment for Medical Classification. In *NeurIPS*, volume 37, 129640–129671.

Kim, E.; Jung, D.; Park, S.; Kim, S.; and Yoon, S. 2023. Probabilistic Concept Bottleneck Models. In *ICML*, 16521–16540. PMLR.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *ICML*, 5338–5348. PMLR.

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

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated graph sequence neural networks. In *ICLR*.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, 15305–15314.

Lin, Y.; Dou, X.; Luo, X.; Wu, Z.; Liu, C.; Luo, T.; Wen, J.; Ling, B. W.-k.; Xu, Y.; and Wang, W. 2025. Multi-view diabetic retinopathy grading via cross-view spatial alignment and adaptive vessel reinforcing. *Pattern Recognition*, 164: 111487.

Liu, C.; Yuanxi, Q.; Xu, Q.; Liu, Y.; Wen, J.; Wang, J.; and Luo, X. 2025. Hierarchical Information Aggregation for Incomplete Multimodal Alzheimer's Disease Diagnosis. In *NeurIPS*.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-to-end multi-task learning with attention. In *CVPR*, 1871–1880.

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

Luo, X.; Pu, Z.; Xu, Y.; Wong, W. K.; Su, J.; Dou, X.; Ye, B.; Hu, J.; and Mou, L. 2021. MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. *Pattern Recognition*, 120: 108104.

Oikarinen, T.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-free Concept Bottleneck Models. In *ICLR*.

Prasse, K.; Knab, P.; Marton, S.; Bartelt, C.; and Keuper, M. 2025. DCBM: Data-Efficient Visual Concept Bottleneck Models. In *ICML*.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. In *NeurIPS*, volume 34, 8583–8595.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2021. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30: 1771–1783.

Wen, C.; Ye, M.; Li, H.; Chen, T.; and Xiao, X. 2024. Concept-based Lesion Aware Transformer for Interpretable Retinal Disease Diagnosis. *IEEE Transactions on Medical Imaging*.

Wilkinson, C. P.; Ferris III, F. L.; Klein, R. E.; Lee, P. P.; Agardh, C. D.; Davis, M.; Dills, D.; Kampik, A.; Pararajasegaram, R.; Verdaguer, J. T.; et al. 2003. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9): 1677–1682.

Xu, Q.; Luo, X.; Huang, C.; Liu, C.; Wen, J.; Wang, J.; and Xu, Y. 2024a. HACDR-Net: Heterogeneous-aware convolutional network for diabetic retinopathy multi-lesion segmentation. In *AAAI*, 6342–6350.

Xu, X.; Qin, Y.; Mi, L.; Wang, H.; and Li, X. 2024b. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *ICLR*.

Yan, S.; Yu, Z.; Zhang, X.; Mahapatra, D.; Chandra, S. S.; Janda, M.; Soyer, P.; and Ge, Z. 2023. Towards trustable skin cancer diagnosis via rewriting model's decision. In *CVPR*, 11568–11577.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 19187–19197.

Yuksekgonul. 2022. Post-hoc Concept Bottleneck Models. In *ICLR*.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc Concept Bottleneck Models. In *ICML*.

Zhang, K.; Liu, X.; Xu, J.; Yuan, J.; Cai, W.; Chen, T.; Wang, K.; Gao, Y.; Nie, S.; Xu, X.; et al. 2021. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nature biomedical engineering*, 5(6): 533–545.