

The 7<sup>th</sup> International Conference on Medical Imaging and Computer-Aided Diagnosis

**Paper ID: 576**

**Title: ORCHID-DR: Clinical Ordinal-Lesion Attention with Semantic Prototypes for Robust Diabetic Retinopathy Grading**

**Authors: Sulaiman Khan, Md. Rafiul Biswas, Zubair Shah\***

# 589 million adults live with diabetes; DR threatens sight



## Diabetes

- **Diabetes** is a chronic condition that reflects the body's inability to produce or respond to insulin.
  - It is characterized by abnormal glucose levels in the blood.
- Diabetes is a large and growing global burden [1].
  - 1 in 9 adults (589 million, aged 20–79) lived with diabetes in 2024; IDF projects 853 million by 2050.
  - Diabetes caused 3.4 million deaths in 2024, one every six seconds.
- Diabetes damages many organs, including the eyes.
  - Complications include heart disease, kidney disease, **diabetic retinopathy (DR)** and osteoporosis.
- This work focuses on **diabetic retinopathy** analysis.



# DR severity follows an ordinal, lesion-driven progression

## Clinical Significance

- **DR severity grading** is crucial for early intervention and preventing irreversible vision loss.
  - Diagnosis rests on identifying distinct pathological **lesion types**.
- DR grades are **ordinal**: severity rises in a fixed order from Grade 0 to Grade 4.
  - So confusing Grade 0 with Grade 4 is far worse than confusing Grade 1 with Grade 2.

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| <b>Grade 0</b> | <b>No DR — Healthy retina without lesions</b>                         |
| <b>Grade 1</b> | <b>Mild — Isolated Microaneurysms</b>                                 |
| <b>Grade 2</b> | <b>Moderate — Hemorrhages &amp; Hard Exudates</b>                     |
| <b>Grade 3</b> | <b>Severe — Dense Hemorrhages &amp; Cotton Wool Spots</b>             |
| <b>Grade 4</b> | <b>PDR (Proliferative DR) — Vision-threatening Neovascularization</b> |

*Next: two problems that stop current models from respecting this order.*

# Class imbalance and grade ambiguity limit current DR models

## ↓ 1. Extreme Class Imbalance

- Screening datasets are heavily **long-tailed**: a few grades hold most of the images.
  - Healthy eyes dominate while vision-threatening stages (Severe / PDR) are sparse.
  - So, a model learns a bias toward the majority No DR class.
- **Impact** – The model misses more critical patients (high false-negative rate).

## 2. Adjacent-Grade Ambiguity

- Standard classifiers treat grades as **unordered classes**.
  - So, stage boundaries between neighbors (e.g., Mild vs. Moderate) blur.
- CNN and transformer models act as **black boxes** [2, 3].
  - They ignore the lesion-based reasoning and severity hierarchy that ophthalmologists use.
- **Impact** – The model can make unsafe distant-grade errors (e.g., PDR graded as No DR).

4

Next: ORCHID-DR, a framework designed to address both problems.

# Research Aim and Contributions

**Aim:** ORCHID-DR (ordinal robust clinical hierarchical lesion-integrated diffusion) predicts the DR grade **autoregressively**: as a chain of yes/no “worse than grade k?” steps, each using the earlier answers (slide 7).

 **COLA Attention** – A learnable clinical importance matrix encodes how each lesion type (e.g., neovascularization) supports each severity grade, as an ophthalmologist weighs evidence.

 **Semantic Prototypes** – Grade-specific prototypes, refined by cross-attention, widen the gap between neighboring DR stages.

 **Adaptive Ordinal-Classification Fusion** – A learnable gate balances cumulative ordinal probabilities with standard multi-class scores (logits).

 **Imbalance-Aware Optimization** – LDAM (label-distribution-aware margin) loss, class-balanced and focal losses (which up-weight rare and hard cases), and minority oversampling protect sensitivity on rare, vision-threatening grades.

5

*Next: how these four components fit together in one architecture.*

# ORCHID-DR mirrors the clinician's lesion-based reasoning

## 1. Backbone

EfficientNet-B3 or ViT-B/16 with multi-scale fusion yields global image features.

## 2. COLA attention

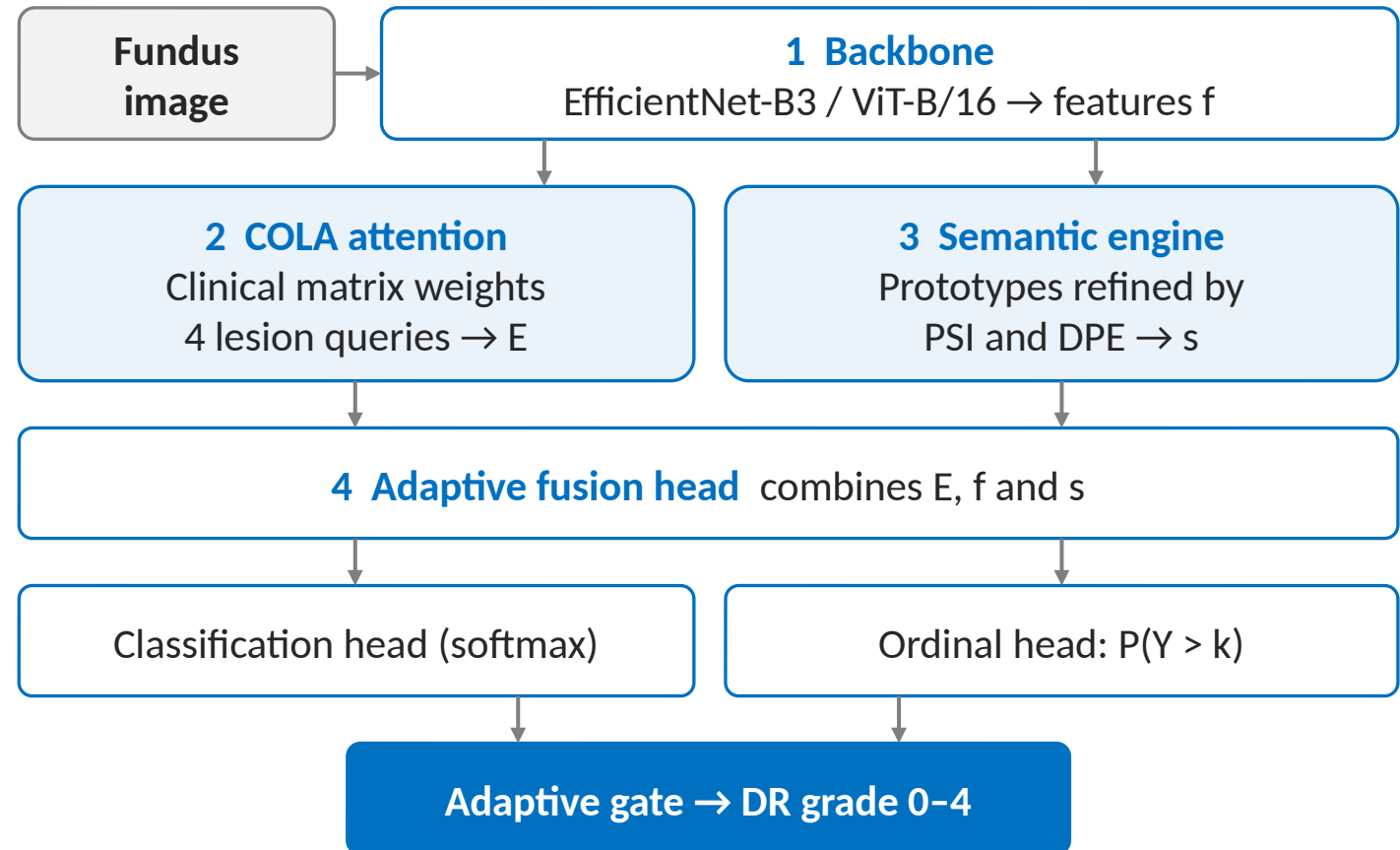
Four lesion queries, weighted by a learnable clinical matrix, gather grade-specific evidence (slide 8).

## 3. Semantic engine

Grade prototypes, refined by two modules (PSI and DPE, slide 9), reduce adjacent-grade confusion.

## 4. Adaptive fusion head

A gate blends ordinal and classification outputs into the final grade (0-4).



# DR grading as an autoregressive sequence of K-1 thresholds

- **Data Labelling** – We reformulate K-class DR grading as a sequence of **K-1 dependent binary objectives**.
  - Each objective asks “is the grade above threshold k?”, so each image becomes a cumulative threshold vector:

$$\text{Class 3 (Severe DR)} \rightarrow y = [1, 1, 1, 0]$$

- Following the **autoregressive** paradigm, each threshold decision uses the image and all earlier decisions.
  - Training maximizes the conditional likelihood of the whole threshold sequence:

$$\max \mathcal{P}(y_1, \dots, y_{K-1}) = \max \prod_{i=1}^{K-1} p(y_i | y_{<i}, I)$$

## Key Advantages

- ✓ **Structural Ordering:** The cumulative encoding keeps every prediction in grade order.
- ✓ **No Distance Collapse:** A distant error is wrong at several thresholds, so the loss penalizes it more.
- ✓ **Probabilistic Softness:** Each threshold outputs a probability, so borderline cases get soft decisions.

7

# COLA weights lesion evidence by its relevance to each grade

- COLA uses four explicit **lesion query types**, each linked to the grades it signals:
  - Microaneurysms** (Mild indicator)
  - Hemorrhages** (Moderate/Severe indicator)
  - Hard Exudates** (Exudative severity indicator)
  - Neovascularization** (PDR critical marker)
- A **learnable clinical matrix** weights how strongly each lesion type supports each grade.
  - Tri-linear attention** combines lesion queries, the grade embedding and image tokens:
 
$$Attn = softmax((QKT / \sqrt{d}) \odot (g k RT(l, k))) V$$
- This mimics how ophthalmologists weigh evidence: some lesions count more for some grades.
  - For example, neovascularization is the critical marker for PDR.

# Grade prototypes pull neighboring DR stages apart

- A **semantic engine** learns one **prototype** per grade: a reference embedding of a typical image of that grade.
- These prototypes are refined via a two-stage modulation process:
  - 1 **Pathological Semantic Injector (PSI)** – Integrates lesion-level descriptors  $s$  using cross-attention to incorporate clinical knowledge into prototypes.
  - 2 **Discriminative Prototype Enhancer (DPE)** – Amplifies inter-grade separation using learnable scaling weights to prevent grade overlap.
- The engine is supervised by a composite semantic loss:

$$\mathcal{L}_{\text{sem}} = \sum_i \|\mathbf{z}_i - \mathbf{p}_{y_i}^*\|_2^2 + \gamma \sum_{k \neq i} \max(0, m - \|\mathbf{p}_k^* - \mathbf{p}_j^*\|_2)$$

- The first term pulls each image embedding toward its own grade's prototype.
- The second term pushes prototypes of different grades at least a margin  $m$  apart.

9

Next: the datasets, baselines and metrics used to test ORCHID-DR.

# Tested on 4 public datasets against 8 state-of-the-art (SOTA) methods

## Datasets

### 4 Public Benchmark datasets

- **APTOS 2019** (High long-tail)
- **Messidor** (Clinical baseline)
- **DDR Dataset** (Large-scale, noisy)
- **DeepDR Challenge** (Multi-site)

## Baselines

### 8 Evaluated SOTA Baselines

ViT-B/16 Baseline

CORAL (Pattern Rec 20)

CORN (PAA 2023)

PoEs (CVPR 2021)

GoL (NeurIPS 2022)

Ord2Seq (ICCV 2023)

CLIP-DR (MICCAI 24)

AOR-DR (MICCAI 25)

**Setup:** EfficientNet-B3 or ViT-B/16 (ImageNet-pretrained), batch 32, up to 30 epochs with early stopping, horizontal-flip augmentation.

**Metrics:** accuracy, F1, sensitivity, specificity and **QWK** (quadratic weighted kappa), which penalizes an error more the further it is from the true grade.

*Next: how ORCHID-DR compares with the eight baselines.*

# ORCHID-DR outperforms 8 SOTA methods on 3 of 4 benchmarks

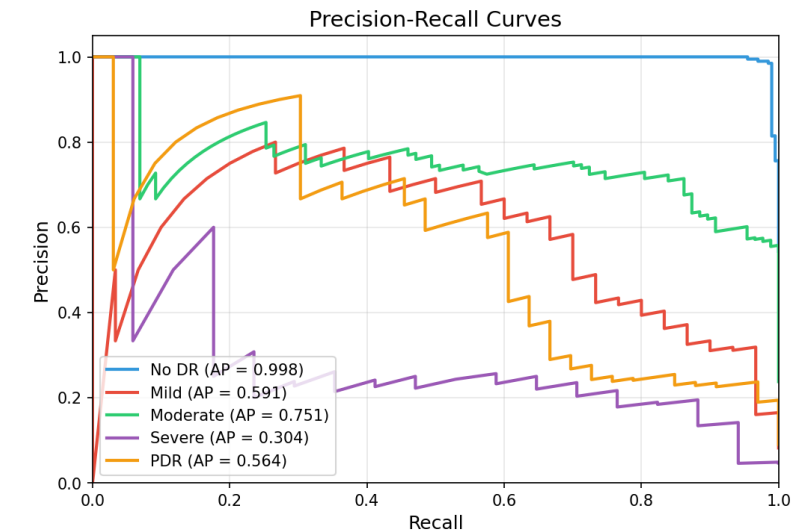
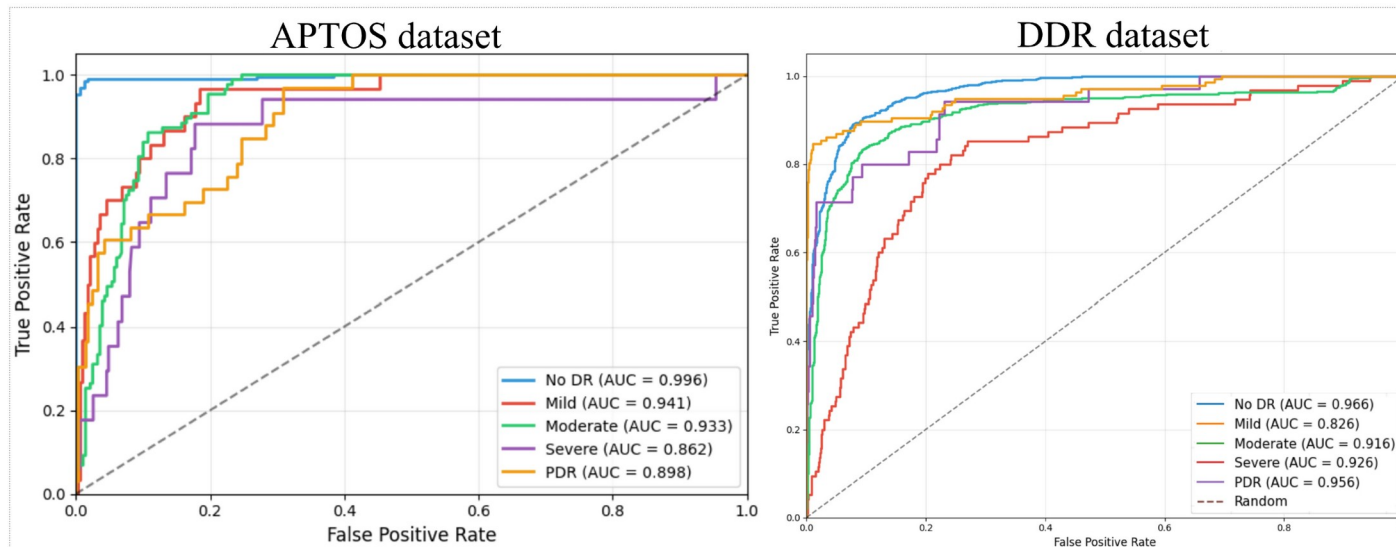
**Table.** Comparison with SOTA ordinal regression models: best result in bold, second best underlined; × = QWK not reported.

| Model               | APTOS       |             |             |             |             | Messidor    |             |             |             |             | DDR         |             |             |             |             | DeepDR      |             |             |             |             |
|---------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                     | Acc         | F1          | Sen         | Spec        | QWK         | Acc         | F1          | Sen         | Spec        | QWK         | Acc         | F1          | Sen         | Spec        | QWK         | Acc         | F1          | Sen         | Spec        | QWK         |
| ViT-B/16            | 69.9        | 38.3        | 36.4        | 68.6        | ×           | <u>63.9</u> | 41.9        | 43.2        | 72.7        | ×           | 57.4        | 26.0        | 22.3        | 63.6        | ×           | 62.5        | 39.1        | 39.8        | 71.7        | ×           |
| CORAL               | 66.4        | 30.4        | 27.3        | 67.8        | ×           | 62.2        | 18.7        | 29.0        | 72.3        | ×           | 54.1        | 26.6        | 24.7        | 67.2        | ×           | 57.7        | 37.3        | 36.9        | 71.2        | ×           |
| CORN                | 69.6        | 31.4        | 27.9        | 60.2        | ×           | 62.2        | 18.6        | 24.9        | 72.3        | ×           | 54.7        | 26.3        | 23.9        | 66.5        | ×           | 56.2        | 31.4        | 28.6        | 71.1        | ×           |
| PoEs                | 67.2        | 32.4        | 30.4        | 68.2        | ×           | 44.7        | 28.1        | 31.8        | 71.2        | ×           | 51.3        | 28.0        | 41.6        | 68.8        | ×           | 55.5        | 31.3        | 34.2        | 71.8        | ×           |
| GoL                 | 74.2        | 52.2        | 60.7        | 71.4        | ×           | 60.8        | 33.8        | 41.4        | 72.6        | ×           | 63.6        | 40.9        | 46.1        | 69.3        | ×           | 62.7        | 51.2        | 56.7        | 72.8        | ×           |
| Ord2Seq             | 70.2        | 30.4        | 27.7        | 64.8        | ×           | <b>64.7</b> | 34.1        | 44.5        | 72.6        | ×           | 64.7        | 34.1        | 36.9        | 65.5        | ×           | 61.2        | 43.2        | 52.4        | 71.3        | ×           |
| CLIP-DR             | 71.8        | 45.5        | 44.5        | 71.1        | ×           | 63.3        | 36.8        | 45.8        | 72.7        | ×           | 68.4        | 45.6        | 50.5        | 69.2        | ×           | 64.5        | 50.6        | 57.1        | 72.5        | ×           |
| AOR-DR              | 77.1        | 58.6        | <u>62.6</u> | 71.9        | ×           | 62.5        | <u>45.1</u> | <b>50.6</b> | 73.0        | ×           | 65.0        | 43.9        | 49.5        | 69.7        | ×           | 65.0        | 51.9        | 57.7        | 72.9        | ×           |
| Ours (ViT-B/16)     | <u>79.5</u> | <u>64.4</u> | <b>65.5</b> | <u>94.8</u> | <u>87.9</u> | 63.7        | <b>46.1</b> | <u>49.3</u> | <b>87.5</b> | <u>70.8</u> | <u>80.2</u> | <u>60.9</u> | <u>59.2</u> | <u>93.9</u> | <u>83.4</u> | <u>66.5</u> | <u>57.0</u> | <u>58.0</u> | <u>91.3</u> | <u>76.3</u> |
| Ours (EfficientNet) | <b>82.8</b> | <b>82.5</b> | 60.2        | <b>95.7</b> | <b>90.3</b> | 57.2        | 42.5        | 46.6        | <u>85.2</u> | <b>77.6</b> | <b>82.5</b> | <b>67.4</b> | <b>67.9</b> | <b>94.9</b> | <b>87.6</b> | <b>70.0</b> | <b>62.7</b> | <b>64.1</b> | <b>92.6</b> | <b>83.0</b> |

**Messidor exception:** Ord2Seq is 1.0 pt higher in accuracy and AOR-DR 1.3 pt higher in sensitivity. Messidor's limited sample size may favor simpler models.

# Per-grade ROC stays strong on rare Severe and PDR grades

- **Rare grades:** ROC curves on APTOS and DDR stay high for the rare Severe and PDR grades (left).
  - An **AUC** of 1.0 means perfect separation; 0.5 means chance.
- **APTOS per grade:** AUC ranges from 0.862 (Severe) to 0.996 (No DR) (left).
- **Precision-recall:** Average precision is lower for rare grades (0.30 Severe, 0.56 PDR) (right).
  - Each false positive weighs more when a grade has few true cases.

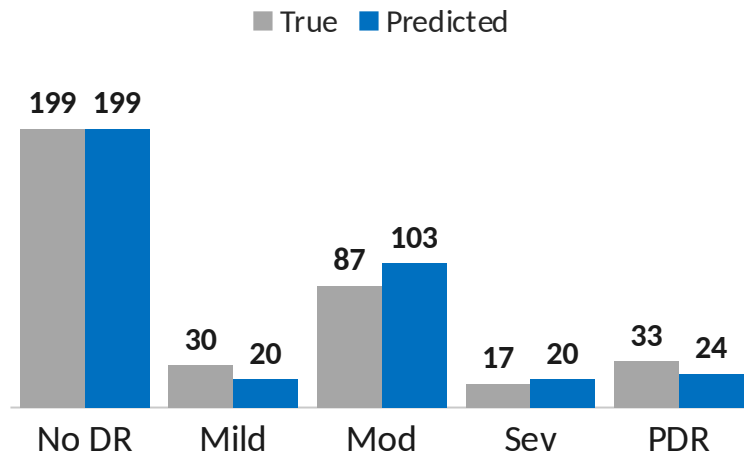


Next: where the remaining errors fall, and whether they are safe.

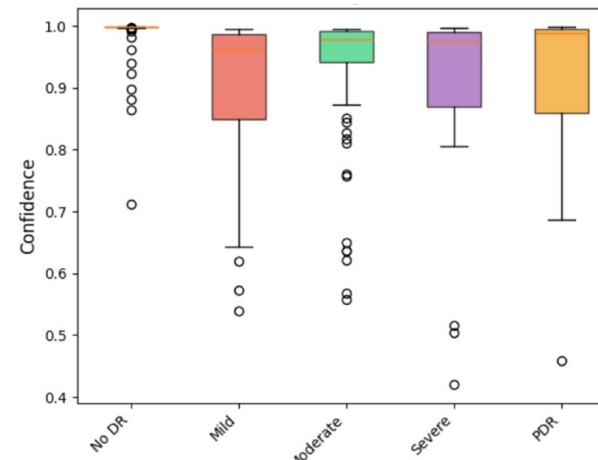
# 303 of 366 APTOS images graded exactly; 75% of errors adjacent

- (a) **Distribution:** The predicted grade counts closely match the ground truth.
- (b) **Confidence:** Median confidence is above 0.95 in every grade; Mild varies most.
- (c) **Errors:** 47 of 63 misclassifications (75%) fall in a neighboring grade.
  - **None of the 50 Severe or PDR images is graded No DR**, so no sight-threatening case is sent home as healthy.

(a) True vs predicted counts



(b) Confidence by true grade



(c) Misclassified images, true × predicted

| True \ Pred | No DR | Mild | Mod | Sev | PDR |
|-------------|-------|------|-----|-----|-----|
| No DR       | ✓     | 2    | 1   | .   | .   |
| Mild        | 1     | ✓    | 13  | 1   | 1   |
| Moderate    | 2     | 3    | ✓   | 6   | 3   |
| Severe      | .     | .    | 9   | ✓   | 4   |
| PDR         | .     | 1    | 7   | 9   | ✓   |

13

Next: does performance hold on datasets the model never saw in training?

# Domain shift lowers QWK, but specificity stays $\geq 85\%$

## 1. Cross-Dataset Generalization Analysis

- **In-domain:** The model is near-optimal when trained and tested on APTOS (QWK 97.1, F1 89.6).
- **Zero-shot transfer:** QWK drops when the model is tested on a dataset it never saw in training.
  - APTOS-trained QWK falls 49.3% on DDR (97.1  $\rightarrow$  49.2).
  - DDR-trained QWK falls only 19.5% on APTOS (87.8  $\rightarrow$  70.7).
- **Takeaway:** Training on the larger, more varied DDR set yields more transferable features.
  - Specificity stays  $\geq 85\%$  in every run, so healthy cases are still identified reliably.

| In-domain                       |      |      |      |      |      | Cross-domain (zero-shot)     |      |      |      |      |      |
|---------------------------------|------|------|------|------|------|------------------------------|------|------|------|------|------|
| Train $\rightarrow$ Test        | Acc  | F1   | Sen  | Spec | QWK  | Train $\rightarrow$ Test     | Acc  | F1   | Sen  | Spec | QWK  |
| DDR $\rightarrow$ DDR           | 82.7 | 67.5 | 67.9 | 95.0 | 87.8 | DDR $\rightarrow$ APTOS      | 48.2 | 40.8 | 39.5 | 87.3 | 70.7 |
| APTOS $\rightarrow$ APTOS       | 93.6 | 89.6 | 91.8 | 98.5 | 97.1 | APTOS $\rightarrow$ DDR      | 58.5 | 37.3 | 38.2 | 85.6 | 49.2 |
| Messidor $\rightarrow$ Messidor | 57.2 | 42.5 | 46.6 | 85.2 | 77.6 | APTOS $\rightarrow$ Messidor | 53.9 | 37.5 | 36.8 | 85.2 | 58.0 |

14

Next: which components drive these results?

# Every module adds accuracy; removing COLA costs the most

## 2. Component Contribution Analysis (DDR Dataset)

- **Full model:** The full model is best on every metric (QWK 83.7 vs 80.4 for the EfficientNet baseline).
- **Without COLA:** Removing COLA causes the largest drop, QWK 83.7  $\rightarrow$  80.8 (−2.9 pts).
  - So weighting lesions by their relevance to each grade contributes the most.
- **Ordinal constraints:** Removing them lowers QWK from 83.7 to 81.5.
  - All ablation runs used 8 epochs, so scores are lower than in the main results table.

Table. Component-wise ablation on the DDR dataset.

| Configuration           | COLA | Semantic | Ordinal | Acc  | F1   | Sen  | Spec | QWK  |
|-------------------------|------|----------|---------|------|------|------|------|------|
| Full model              | ✓    | ✓        | ✓       | 78.7 | 64.6 | 64.8 | 93.3 | 83.7 |
| w/o COLA                | ×    | ✓        | ✓       | 74.9 | 60.4 | 61.0 | 92.9 | 80.8 |
| w/o Semantic            | ✓    | ×        | ✓       | 73.7 | 62.9 | 64.7 | 92.8 | 81.4 |
| w/o Ordinal             | ✓    | ✓        | ×       | 74.5 | 61.6 | 64.7 | 92.9 | 81.5 |
| w/o COLA + Semantic     | ×    | ×        | ✓       | 73.2 | 58.7 | 60.5 | 88.3 | 80.6 |
| Baseline (EfficientNet) | ×    | ×        | ×       | 73.9 | 59.2 | 61.0 | 90.9 | 80.4 |

15

Next: what these results mean for automated DR screening.

# Conclusion



**Clinical Alignment** – ORCHID-DR builds lesion-based, ordinal ophthalmic reasoning into the network itself.



**Safety on Severe Cases** – No Severe or PDR image was graded as healthy on APTOS.



**Cross-Dataset Specificity** – Specificity stayed  $\geq 85\%$  when the model was tested on datasets it never saw.



**Every Module Contributes** – Removing any one module lowers QWK below the full model's 83.7.

# References

- [1]. International Diabetes Federation, *IDF Diabetes Atlas*, 11th ed. Brussels: IDF, 2025. [diabetesatlas.org](https://diabetesatlas.org)
- [2]. Khan, S., Biswas, M.R., Shah, Z.: DiabEye-Q: AI-driven longitudinal analysis of ophthalmoscopic images for early diabetes prediction in Qatari adults. *In: International Conference on AI in Healthcare*, pp. 293–306. Springer (2025).
- [3]. Saleem, S., Talib, R., Hanif, M. K., & Sarwar, M. U.: Clinical-guided deep learning framework for diabetic retinopathy: integrating lesion-aware attention, adversarial augmentation, and uncertainty quantification. *Scientific Reports* (2026).

# Questions & Discussion

## Author Contacts:

Sulaiman Khan: [sukhan@hbku.edu.qa](mailto:sukhan@hbku.edu.qa)

Md Rafiul Biswas: [mbiswas@hbku.edu.qa](mailto:mbiswas@hbku.edu.qa)

Zubair Shah\*: [zshah@hbku.edu.qa](mailto:zshah@hbku.edu.qa)