Horizon-Aware Vision–Language Forecasting of Diabetic Retinopathy with Text Prototypes

Gongyu Zhang[1,†] **Yutong Li**[2,†] **Zengxiang Li**[3] **Timothy Jackson**[1,*] **Christos Bergeles**[1,*]
[1]King's College London [2]University College London [3]Singhealth [†, *] Equal contribution
gongyu.zhang@kcl.ac.uk, timothy.jackson@kcl.ac.uk, christos.bergeles@kcl.ac.uk

Abstract

Identifying eyes at high risk of future diabetic retinopathy (DR) is valuable for recall scheduling and timely intervention. We present a vision–language forecasting framework that *supervises a contrastive alignment* between fundus images and *horizon-aware* hypothesis prompts using a *multi-positive, bidirectional contrastive objective* with *same-image hard negatives*. At inference, we classify by *text prototypes*: per-class prompt sets are encoded and averaged to prototypes, and image–prototype similarities yield calibrated risk. On a national screening cohort, adding simple demographic context (age, sex, laterality) *inside the prompts* improves forecasting over image-only baselines across 1/2/3-year horizons (e.g., 1-year AUROC $0.654 \rightarrow 0.673$ with age; 0.683 with age+sex). Our results establish a compact, label-efficient VLM baseline for multi-horizon DR risk that keeps language grounding and supports prototype-based classification and post-hoc calibration.

1 Introduction

Diabetic retinopathy (DR), a microvascular complication of diabetes and a leading cause of preventable blindness worldwide, demands timely intervention based on accurate risk stratification American Diabetes Association Professional Practice Committee [2024]. Screening programmes routinely collect retinal fundus photographs, enabling automated detection of referable DR (*e.g.*, R2+ or M1), and deep learning systems have demonstrated robust performance for contemporaneous diagnosis Gulshan et al. [2016], Ting et al. [2017], Abràmoff et al. [2018]. However, for clinical workflows—ranging from optimising recall intervals to resource allocation and triage—it is often more valuable to forecast whether an eye will progress within specific time horizons (e.g., one, two, or three years). In this work we treat forecasting as a first-class objective and ask whether multimodal alignment between images and explicit, horizon-aware hypotheses can improve risk estimates without sacrificing language grounding or extensibility.

Emerging evidence indicates that retinal images encode latent prognostic information. Poplin et al. showed that fundus photos can predict cardiovascular risk factors via deep learning Poplin et al. [2018], and subsequent work extended this idea to multi-year DR forecasting Bora et al. [2021], Rom et al. [2022]. More recently, DeepDR Plus modelled time-to-progression to personalise screening schedules Dai et al. [2024]. Yet these efforts largely remain within image-only frameworks and seldom leverage structured patient context or hypothesis-level supervision. In particular, they do not exploit the natural compatibility between future-oriented clinical questions ("remain healthy" vs. "progress to referable DR" within a given horizon) and language-conditioned decision rules that can be calibrated and adapted post hoc.

Within ophthalmology, initiatives such as RETFound, FLAIR, and EyeCLIP pretrain on large retinal corpora or encode expert language supervision, improving contemporaneous classification and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

description tasks Zhou et al. [2023], Silva-Rodríguez et al. [2025], Shi et al. [2025]. Nevertheless, most existing VLMs are optimised for same-time recognition rather than forecasting future disease, and few incorporate horizon-aware hypotheses, multi-positive supervision from paraphrases, or explicit treatment of *same-image* negatives arising from discrepant labels or horizons.

Gap and opportunity. Horizon-aware vision–language forecasting remains underexplored. We posit that future-risk prediction can be cast as aligning fundus images with concise, structured prompts that encode demographics and laterality together with outcome hypotheses at specified horizons. This framing enables supervision via a *multi-positive* contrastive objective (to exploit paraphrase diversity) while emphasising *same-image hard negatives* (same eye, different label/horizon) to sharpen decision boundaries—retaining language grounding and zero/low-shot extensibility.

Our Contribution. We introduce a compact VLM for multi-horizon DR forecasting that (i) pairs retinal images with *horizon-aware* hypothesis prompts containing simple demographics; (ii) trains using a *supervised, multi-positive bidirectional contrastive objective* with *same-image hard negatives*; and (iii) performs classification via *text prototypes* built from multiple paraphrases per class, enabling calibrated, prompt-driven inference. On a national screening cohort, injecting demographics into prompts consistently improves over image-only baselines across 1/2/3-year horizons (e.g., at 1-year, AUROC $0.654 \rightarrow 0.673$ with age; 0.683 with age+sex), establishing a strong, reproducible baseline for multimodal DR risk forecasting.

2 Methods

2.1 Data and outcomes

We curate a development cohort of 27,863 patients (55,533 eyes; 403,951 images) from a subset of the UK national diabetic eye screening programme Nderitu et al. [2022] and hold out 11,198 eyes for internal testing. Patients have sufficient longitudinal follow-up to define horizon outcomes. Each eye–session provides two-field colour fundus photographs (macula and disc) plus metadata: age, sex, and laterality (L/R). For horizons at 1, 2, and 3 years we create binary labels indicating whether the eye becomes *referable* within the horizon (referable DR, e.g., R2+, or maculopathy, e.g., M1) versus non-referable. We enforce *patient-level* splits—both eyes and all visits of a patient remain in the same partition—to prevent leakage. Unless noted, the macular field is used for training/inference; the disc field is retained for sensitivity checks. Mini-batches are approximately horizon- and label-balanced to stabilise optimisation and calibration.

2.2 Model and supervised contrastive training

Clinical context is rendered as concise, horizon-aware hypotheses with minimal demographics; e.g., *"62-year-old male; left eye. Will the eye remain healthy within 1 year?"* and its complement "*. . . develop referable DR within 1 year.*" Multiple paraphrases are prepared per class/horizon and one is sampled at training time.

Encoders and notation. A vision encoder (e.g., ViT-B/16) maps an image to $v \in \mathbb{R}^d$; a text encoder (e.g., clinical BERT with an optional projector) maps a prompt to $t \in \mathbb{R}^d$. Let $\widehat{v} = v/\|v\|_2$, $\widehat{t} = t/\|t\|_2$. A learnable temperature $\tau = \exp(\theta) > 0$ scales cosine similarities. For a batch of size B,

$$S \in \mathbb{R}^{B\times B}, \qquad S_{ij} = \tau \,\langle \widehat{v}_i, \widehat{t}_j \rangle.$$

Positive groups and soft targets. With horizon $h \in \{1,2,3\}$ and label $y \in \{\mathrm{H}, \mathrm{R}\}$ (**H**=*healthy within horizon*, **R**=*referable*), define

$$\texttt{pos\_id} = (\texttt{image\_uid},\, h,\, y).$$

Different paraphrases sharing a pos_id are positives. Build $Y \in \{0,1\}^{B\times B}$ with $Y_{ij} = 1$ iff $\texttt{pos\_id}(v_i) = \texttt{pos\_id}(t_j)$, and normalise to soft targets

$$T^{\text{row}}_{i,:} = \frac{Y_{i,:}}{\sum_k Y_{ik} + \varepsilon}, \qquad T^{\text{col}}_{:,j} = \frac{Y_{:,j}}{\sum_k Y_{kj} + \varepsilon}.$$

Bidirectional contrastive loss and hard negatives.

$$\mathcal{L}_{\mathrm{i}\to\mathrm{t}} = -\frac{1}{B}\sum_{i}\sum_{j} T_{ij}^{\mathrm{row}} \log \frac{\exp(S_{ij})}{\sum_k \exp(S_{ik})}, \qquad \mathcal{L}_{\mathrm{t}\to\mathrm{i}} = -\frac{1}{B}\sum_{j}\sum_{i} T_{ij}^{\mathrm{col}} \log \frac{\exp(S_{ij})}{\sum_k \exp(S_{kj})}, \tag{1}$$

and $\mathcal{L} = \mathcal{L}_{\mathrm{i}\to\mathrm{t}} + \mathcal{L}_{\mathrm{t}\to\mathrm{i}}$. To stress *same-image hard negatives*—same image but mismatched (h, y)—let

$$\mathcal{H}_i = \{\, j : \texttt{image\_uid}(t_j) = \texttt{image\_uid}(v_i) \text{ and } (h, y) \text{ mismatch} \,\},$$

and upweight their contributions in the softmax denominators by $\lambda = 2.0$ (equivalently add $\log \lambda$ to such logits). We train with AdamW (lr 1×10^{-4}), weight decay 1×10^{-5}, cosine decay with one-epoch warm-up, standard augmentations, and jointly learn τ.

2.3 Prototype-based inference and calibration

For each class–horizon $c \in \{\mathrm{H@1y}, \mathrm{R@1y}, \mathrm{H@2y}, \mathrm{R@2y}, \mathrm{H@3y}, \mathrm{R@3y}\}$, encode K paraphrases $\{t_{c,k}\}_{k=1}^{K}$ and form the normalised prototype

$$p_c = \frac{1}{K}\sum_{k=1}^{K} t_{c,k}, \quad p_c \leftarrow \frac{p_c}{\|p_c\|_2}.$$

Given an image embedding v, logits are $s_c = \tau\,\langle \widehat{v}, p_c \rangle$. We report (i) binary H vs. R at a fixed horizon, or (ii) a multi-class decision over all six prototypes. Let $P@k$ denote the probability of *referable within k years*. We apply temperature calibration on validation data and optionally enforce monotonicity $P@3 \geq P@2 \geq P@1$ (e.g., isotonic regression), yielding calibrated, horizon-consistent risks while retaining the interpretability of prototype editing.

Statistical analysis and evaluation protocol. We report AUROC, F1, accuracy, sensitivity, and specificity on the held-out test set using thresholds selected on validation data.

3 Results

3.1 Forecasting performance across horizons

Table 1 summarises discrimination and operating characteristics across 1-, 2-, and 3-year horizons. Using the image-only configuration as the reference ("Rem/Dev-*y" rows), injecting demographic context into the horizon-aware prompts improves AUROC at every horizon while preserving the same contrastive setup. At 1 year the baseline AUROC of 0.654 rises to a best of 0.691 when *eye+sex+age* are included, with simpler variants already effective (e.g., *age* 0.673, *age+eye* 0.685). At 2 years AUROC increases from 0.656 to 0.686 with *age+eye*, with *sex+age* close behind at 0.684. At 3 years AUROC improves from 0.649 to 0.671 with *eye+sex+age*. These gains are mirrored by F1 and accuracy. For 1 year, the *age* variant attains the highest F1 (0.428) and *eye+sex+age* yields the best accuracy (0.631). For 2 years, *age* yields the highest F1 (0.419) while *age+eye* provides the strongest accuracy (0.628). For 3 years, *age+eye* maximises F1 (0.389) and *sex+age* slightly edges accuracy (0.611). Beyond aggregate discrimination, demographic-aware prompts generally increase sensitivity with modest specificity trade-offs. At 1 year, for example, sensitivity increases from 0.602 to 0.725 under *eye+sex+age*, with specificity moving from 0.562 to 0.535; analogous patterns hold at 2 years (*age+eye* sensitivity 0.651 vs. 0.601 baseline) and at 3 years (*age+eye* 0.622 vs. 0.612 baseline). In screening regimes that prioritise recall, these operating points may be preferable, and post-hoc threshold calibration can recover specificity as required.

3.2 Effect of visit history

Table 2 examines adding limited prior visits as structured context within the same framework. The effect is horizon dependent and most pronounced at 1 year, where AUROC improves from 0.691 (demographics only) to 0.705 with one prior visit and 0.707 with two prior visits; F1 increases from 0.365 to 0.407 and 0.418, and accuracy moves from 0.631 to 0.639 and 0.638. At 2 years, the corresponding AUROC gains are smaller but consistent ($0.674 \rightarrow 0.680 \rightarrow 0.681$) with accuracy

Table 1: Forecasting results grouped by horizon. Demographics are injected via prompts. *Notation:* “Rem/Dev-ky” denotes image-only prompts (no demographics) that contrast *Remain healthy within k years* vs. *Develop referable DR (or maculopathy) within k years*, for $k \in \{1, 2, 3\}$. Suffixes “eye/sex/age” indicate the demographic fields added to the prompt.

Year	AUC	F1	Accuracy	Sensitivity	Specificity	Notes
3-Year	0.649	0.358	0.598	0.612	0.521	Rem/Dev-3y
	0.656	0.367	0.603	0.584	0.560	eye + Rem/Dev-3y
	0.650	0.370	0.600	0.610	0.549	sex + Rem/Dev-3y
	0.656	0.375	0.606	0.592	0.541	age + Rem/Dev-3y
	0.656	0.343	0.594	0.577	0.571	eye + sex + Rem/Dev-3y
	0.657	0.389	0.609	0.622	0.515	age+eye + Rem/Dev-3y
	0.659	0.385	0.611	0.600	0.548	sex + age + Rem/Dev-3y
	0.671	0.382	0.610	0.595	0.543	eye + sex + age + Rem/Dev-3y
2-Year	0.656	0.370	0.603	0.601	0.572	Rem/Dev-2y
	0.666	0.383	0.609	0.595	0.561	eye + Rem/Dev-2y
	0.669	0.406	0.618	0.632	0.548	sex + Rem/Dev-2y
	0.674	0.419	0.624	0.640	0.559	age + Rem/Dev-2y
	0.669	0.385	0.615	0.590	0.583	eye + sex + Rem/Dev-2y
	0.686	0.397	0.628	0.651	0.542	age+eye + Rem/Dev-2y
	0.684	0.401	0.621	0.633	0.557	sex + age + Rem/Dev-2y
	0.674	0.357	0.611	0.609	0.591	eye + sex + age + Rem/Dev-2y
1-Year	0.654	0.363	0.608	0.602	0.562	Rem/Dev-1y
	0.669	0.354	0.616	0.585	0.571	eye + Rem/Dev-1y
	0.667	0.394	0.624	0.648	0.540	sex + Rem/Dev-1y
	0.673	0.428	0.630	0.655	0.546	age + Rem/Dev-1y
	0.668	0.374	0.618	0.598	0.569	eye + sex + Rem/Dev-1y
	0.685	0.387	0.626	0.644	0.554	age+eye + Rem/Dev-1y
	0.683	0.403	0.629	0.637	0.561	sex + age + Rem/Dev-1y
	0.691	0.365	0.631	0.725	0.535	eye + sex + age + Rem/Dev-1y

Table 2: Performance vs. history length across horizons.

Horizon	History length	AUC	F1	Accuracy	Sensitivity	Specificity
3-year	Demographics only	0.671	0.382	0.610	0.595	0.543
	+1 prior visit	0.674	0.384	0.623	0.608	0.515
	+2 prior visits	0.676	0.362	0.616	0.622	0.609
2-year	Demographics only	0.674	0.357	0.611	0.609	0.591
	+1 prior visit	0.680	0.402	0.624	0.603	0.522
	+2 prior visits	0.681	0.398	0.625	0.625	0.573
1-year	Demographics only	0.691	0.365	0.631	0.725	0.535
	+1 prior visit	0.705	0.407	0.639	0.653	0.642
	+2 prior visits	0.707	0.418	0.638	0.647	0.581

improving to 0.624 and 0.625; F1 rises from 0.357 to 0.402 and remains at 0.398. At 3 years, the AUROC changes are modest ($0.671 \rightarrow 0.674 \rightarrow 0.676$) and the F1 pattern suggests diminishing returns with longer histories ($0.382 \rightarrow 0.384 \rightarrow 0.362$), while accuracy still exceeds the demographics-only setting for one additional visit (0.623 vs. 0.610). Overall, recent history offers the clearest benefit for short-horizon forecasts, with smaller incremental value as the horizon lengthens.

4 Discussion

4.1 Key findings and practical implications

Encoding horizon-aware hypotheses with lightweight demographics directly in the prompts improves multi-horizon DR forecasting while keeping training compact. Limited visit history helps most at 1 year and yields smaller, consistent gains at longer horizons, suggesting recent context is most informative for short-term risk. A supervised, multi-positive bidirectional contrastive objective aligns images with language by using paraphrase diversity as multiple positives and treating discrepant labels or horizons from the same image as hard negatives. After alignment, prototype-based inference is simple and stable: averaging paraphrases forms robust class prototypes, image–prototype scores calibrate well with a single validation temperature, and the same machinery delivers probabilities for 1, 2, and 3 years without task-specific heads. The gains show a typical trade-off, with sensitivity rising more than specificity—appropriate for recall-focused screening—while specificity can be recovered via threshold selection or post-hoc calibration. Practically, balanced batches and paraphrase sampling surface multi-positive structure; including age (and, when available, laterality) preserves language grounding; and optional monotonic post-processing enforces clinically sensible ordering across horizons without altering the training loss.

4.2 From discrete horizons to survival-style inference

Although we report results at three fixed horizons, the same prototype workflow can be extended to time-to-event prediction. Instead of asking "progress within one, two, or three years," we can build a small library of prompts that describe risk month-by-month or visit-by-visit. The model then produces a sequence of per-interval risks that naturally form a survival curve and its running incidence, giving clinicians a continuous view of how risk accumulates over time. Training can keep the same contrastive objective and add a light regulariser that encourages sensible ordering across time without over-constraining representations. Routine challenges in survival analysis—such as variable follow-up and censoring—can be handled with simple weighting schemes or discrete-time likelihoods while preserving label efficiency. Crucially, the text-prototype design makes the system maintainable: programmes can adjust interval granularity, swap wording, or localise phrases without retraining encoders, and calibration over time can be updated on a small validation set as follow-up policies evolve.

4.3 Limitations and future directions

We train with a single contrastive objective and intentionally avoid auxiliary classification heads, leaving open whether small CE/BCE heads might improve calibration under cost-sensitive operating points. We do not model predictive uncertainty; conformal or Bayesian add-ons could support difficult cases and subgroup analyses. We lack external-site validation, so robustness to acquisition shifts, grading protocols, and demographics remains to be assessed; future work should test cross-programme generalisation and adapt prompts or prototypes accordingly. Our emphasis on same-image hard negatives assumes sufficient within-eye variability in labels or horizons appearing in batches; more systematic batch construction or curriculum strategies could strengthen this signal under extreme imbalance. Finally, richer longitudinal information—structured grades, treatment history, and imaging-derived biomarkers—could be injected into prompts or encoded temporally while retaining the same contrastive training, and calibration could use shared monotonic mappings across horizons to stabilise longitudinal decisions.

5 Conclusion

We introduce a compact, supervised vision–language framework for multi-horizon DR forecasting. A label-supervised, *multi-positive bidirectional contrastive objective* aligns images and horizon-aware hypotheses; classification then uses text prototypes with simple calibration. Demographic-aware prompts consistently improve over image-only baselines, offering a strong, reproducible baseline for multimodal DR risk forecasting.

References

American Diabetes Association Professional Practice Committee. Retinopathy, neuropathy, and foot care: Standards of care in diabetes—2024. *Diabetes Care*, 47(Supplement 1):S231–S247, 2024.

Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016.

Daniel SW Ting, Carol Y Cheung, Gavin Lim, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017.

Michael D Abràmoff, Y Lou, A Erginay, et al. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1(1):39, 2018.

Ryan Poplin, Avinash V Varadarajan, Kay Blumer, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

Abhijit A Bora, Daniel SW Ting, Pearse A Keane, et al. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health*, 3(7):e476–e485, 2021.

Daniel Rom, Elad Dok, Amit Barak, et al. Predicting the future development of diabetic retinopathy using a machine learning approach. *BMJ Open Ophthalmology*, 7(1):e001028, 2022.

L Dai, X Xu, R Liu, et al. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*, 30(4):584–594, 2024.

Y Zhou, K Yu, J He, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 620(7972):123–130, 2023.

J Silva-Rodríguez, H Chakor, R Kobbi, J Dolz, and IB Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99: 103357, 2025.

D Shi et al. A multimodal visual–language foundation model for ocular imaging and ophthalmic diagnosis (eyeclip). *npj Digital Medicine*, 8:–, 2025.

Paul Nderitu, Joan M. Nunez do Rio, Laura Webster, Samantha S. Mann, David Hopkins, M. Jorge Cardoso, Marc Modat, Christos Bergeles, and Timothy L. Jackson. Automated image curation in diabetic retinopathy screening using deep learning. *Scientific Reports*, 12:11196, 2022. doi: 10.1038/s41598-022-15491-1. URL `https://www.nature.com/articles/s41598-022-15491-1`.