A Time-Series Vision–Language Model for Predicting Progression of Diabetic Retinopathy

Abstract

Early detection of diabetic retinopathy (DR) progression is critical for timely intervention and prevention of vision loss. We present a time-series vision–language model that integrates longitudinal clinical context with retinal fundus images to forecast progression to referable DR at 1-, 2-, and 3-year horizons. The framework aligns fundus photographs with structured narrative prompts that encode demographics, diabetes history, and prior screening outcomes. Training is formulated as a contrastive objective, encouraging image embeddings to align with the correct horizon-specific outcome hypothesis. Using a diabetic screening dataset, we show that incorporating longitudinal information into the prompts consistently improves predictive performance, with the best one-year configuration achieving an AUROC of 0.707. The approach offers two key advantages: interpretability, by conditioning predictions on explicit clinical narratives, and extensibility, by allowing prompts to be adapted or enriched with additional timepoint information. To our knowledge, this is the first vision–language framework for horizon-specific DR forecasting, establishing a simple and reproducible baseline for adaptive recall scheduling, triage, and population-level risk management in DR screening programmes.

1 Introduction

Diabetic retinopathy (DR) is a leading cause of preventable blindness worldwide, and its prevalence continues to rise with the increasing burden of diabetes. Screening programs routinely acquire retinal fundus photographs for early detection, and deep learning systems have achieved high accuracy for referable DR detection from single images Gulshan et al. [2016], Ting et al. [2017]. Regulatory milestones, including FDA authorization of an autonomous AI diagnostic system, demonstrate that image-only models can achieve clinical-grade performance Abràmoff et al. [2018]. Yet most current systems focus on contemporaneous diagnosis, whereas clinicians and screening programs often require horizon-specific forecasts—whether an eye will progress within one to three years—to optimize recall intervals and allocate resources World Health Organization, Regional Office for Europe [2020], American Diabetes Association Professional Practice Committee [2024].

Evidence suggests fundus images contain prognostic signals beyond contemporaneous grading. Poplin *et al.* showed that deep learning can infer cardiovascular risk factors from retinal photographs Poplin et al. [2018]. Building on this, several groups proposed image-based forecasting of DR progression, including prediction of incident DR from baseline images Bora et al. [2021], multi-year forecasting Rom et al. [2022], and time-to-progression modeling for personalized screening intervals Dai et al. [2024]. Nderitu *et al.* compared image-only, tabular, and multimodal approaches, highlighting the value of combining clinical context with imaging Nderitu et al. [2024]. These studies confirm feasibility but remain primarily image-centric and rarely exploit language modeling for longitudinal context.

Meanwhile, vision–language models (VLMs) have transformed medical AI. Contrastive pretraining on paired images and reports, as in ConVIRT Zhang et al. [2020], GLoRIA Huang et al. [2021], and BioViL Boecking et al. [2022], yields label-efficient representations competitive with ImageNet

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

pretraining. Larger biomedical VLMs such as MedCLIP, BiomedCLIP, and CheXzero scale corpora and achieve state-of-the-art performance in radiology Wang et al. [2022], Zhang et al. [2023], Tiu et al. [2022]. Despite these advances, most medical VLMs address present-time tasks such as labeling findings or answering visual questions, rather than forecasting future disease states.

Ophthalmology has begun to adopt both unimodal foundation models and retina-specific VLMs. RETFound pretrains a ViT on 1.6M unlabelled retinal images and transfers effectively across diverse tasks Zhou et al. [2023]. FLAIR introduces expert-informed prompts for fundus understanding Silva-Rodríguez et al. [2025], while RetiZero and EyeCLIP scale retina-specific image–text pretraining for broad disease identification Wang and et al. [2025], Shi and et al. [2025]. Yet these models, like their radiology counterparts, are designed for contemporaneous diagnosis rather than multi-horizon forecasting.

This study addresses that gap by proposing a time-series vision–language model for DR risk prediction. Our contributions are threefold: **1.** We present the first vision–language framework for forecasting diabetic retinopathy progression across 1-, 2-, and 3-year horizons. **2.** We introduce a novel time-series VLM that integrates fundus images with structured longitudinal clinical prompts and trains using a symmetric contrastive objective. **3.** We demonstrate, on a large screening dataset, that incorporating longitudinal information consistently improves predictive performance and establish a reproducible baseline for adaptive recall scheduling and population-level risk management.

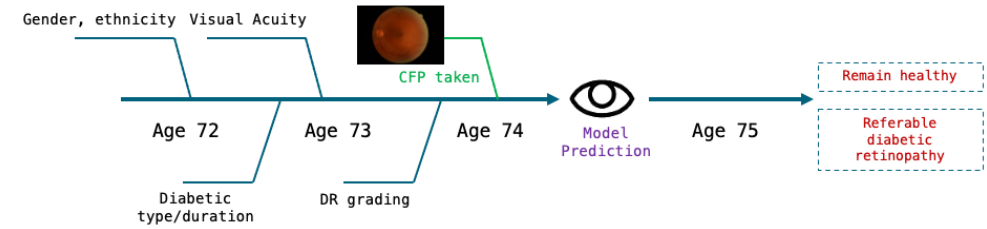
2 Methods

2.1 Dataset

We used a development cohort of 20,900 patients (41,650 eyes) with 302,950 images drawn from a diabetic eye screening programme, of which all patients had sufficient longitudinal follow-up to construct horizon outcomes. Independent evaluation was performed on one held-out internal test set comprising 8,400 eyes from the same regional programme. Each eye–session consisted of a two-field color fundus photograph paired with structured clinical metadata, including routinely captured risk factors such as patient age, sex, duration of diabetes, and baseline DR/Maculopathy status. In this study, we only used the macular images. For each of the three forecast horizons (1, 2, and 3 years), binary outcome labels were constructed indicating whether the patient was referable. A patient was considered referable if either one of the conditions below was met: referable DR (e.g., R2+) or maculopathy (e.g., M1). Otherwise, it was categorized as non-referable. We applied strict patient-level partitioning across development and internal test splits to ensure no patient appeared in both sets. This dataset supports robust evaluation of image-only, tabular-only, and multimodal models across multi-horizon prognostic tasks.

2.2 Text Prompt Construction

We designed a structured narrative prompt to encode patient demographics, diabetes history, and session-level metadata alongside hypothesized outcomes. The base template followed a fixed schema. To increase linguistic diversity and reduce overfitting to a single phrasing, we employed ChatGPT-4o to automatically generate 20 semantically equivalent variations of each prompt template, differing in word order, synonyms, and sentence style. At training time, a random version was sampled for each image–text pair, ensuring exposure to diverse formulations. In addition, a random seed determined how many years of longitudinal information were included, simulating real-world variability in follow-up histories and encouraging robustness to incomplete temporal records. If longitudinal information from prior screening years was available, additional sentences were appended to describe the best visual acuity and retinopathy grade observed at the most recent prior visit. This allowed the model to incorporate temporal dynamics of disease progression. Figure 1 illustrates how longitudinal demographic, clinical, and ophthalmic features are aligned along a patient timeline and encoded into a structured narrative prompt paired with the current fundus image. Each image was paired with two alternative prompts: one ending with a hypothesis that the eye would *remain healthy* within the prediction interval, and the other ending with a hypothesis that the eye would *develop referable diabetic retinopathy*. Only one of the pair corresponded to the ground-truth label, allowing the training objective to be framed as contrastive alignment between images and the correct outcome hypothesis.



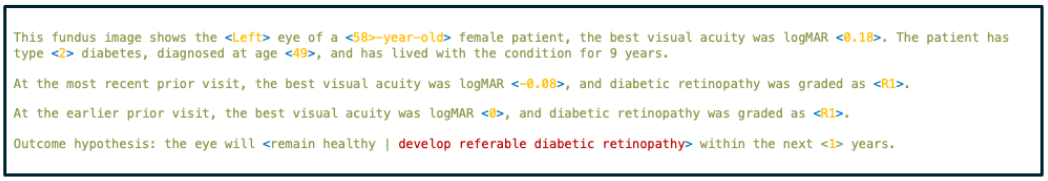


Figure 1: Illustration of the time-series prompt construction. Longitudinal demographic, clinical, and ophthalmic features (e.g., diabetes type, duration, visual acuity, and DR grading) are aligned along a patient timeline. These elements are then encoded into a structured narrative prompt, which is paired with the current fundus image and used by the model to predict future progression to referable diabetic retinopathy.

2.3 Training Strategy

We trained our framework end-to-end to align fundus photographs with structured narrative prompts in a shared embedding space. A Vision Transformer (ViT-B/16) served as the image encoder, while medical pretrained encoders were used as the text encoder. Both modalities were projected into a d-dimensional space and optimized jointly.

Symmetric contrastive alignment. Training followed a contrastive learning setup, where each image was aligned with its horizon-specific outcome prompt and contrasted against negatives from alternative hypotheses and other patients. This yielded a multimodal embedding space suitable for horizon-specific classification.

Temporal encoding. To incorporate longitudinal information, prompts concatenated structured summaries of up to K previous visits along the patient timeline. Let $\mathcal{H}_i = \{h_i^{(1)}, h_i^{(2)}, \ldots, h_i^{(K)}\}$ denote historical descriptors for patient i (e.g., age, diabetes duration, prior DR grade). Each descriptor is mapped to an embedding $e(h_i^{(k)})$, and the temporal context vector is

$$c_i = \frac{1}{K} \sum_{k=1}^{K} e(h_i^{(k)}),$$

which is then injected into the narrative prompt paired with the current fundus image. This averaging formulation acts as a simple temporal memory that captures disease trajectory without requiring explicit interval modeling.

During training, K was sampled randomly from $\{0, 1, 2\}$ to mimic variable follow-up lengths and encourage robustness to incomplete histories. Predictions are thus conditioned not only on the current fundus image but also on an aggregated summary of prior visits, making the model sensitive to temporal disease progression.

Inference. At test time, predictions are obtained by computing cosine similarities between the fundus embedding v_i and the two horizon-specific hypothesis prompts augmented with c_i. A softmax over these similarity scores produces probabilities for *remain healthy* versus *progress*, with the higher-probability class returned as the model's prediction.

Table 1: Prediction performance versus history length across different 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

3 Results and Discussion

3.1 Overall Performance

The time-series vision–language model achieved consistent discrimination for forecasting referable DR across clinically relevant intervals. For the one-year horizon, the best configuration attained an AUROC of 0.707, an F1-score of 0.418, and an accuracy of 0.638 when two prior visits were included in the prompts. For the two-year horizon, the corresponding values were AUROC 0.680, F1-score 0.402, and accuracy 0.624. For the three-year horizon, performance decreased slightly, reaching AUROC 0.676, F1-score 0.362, and accuracy 0.616. As expected, shorter horizons yielded stronger results, reflecting the greater challenge of predicting longer-term outcomes. Across horizons, sensitivity was generally higher than specificity, indicating that the model prioritizes recall of true progressors at the cost of more false positives, which is often acceptable in the context of population-level screening.

Across all horizons, performance improved when prior visit information was incorporated into the text prompts. At one year, adding two prior visits increased AUROC from 0.691 to 0.707, F1-score from 0.365 to 0.418, and accuracy from 0.631 to 0.638. At two years, AUROC rose from 0.674 to 0.680, F1-score from 0.357 to 0.402, and accuracy from 0.611 to 0.624. At three years, AUROC increased from 0.671 to 0.676, F1-score changed from 0.382 to 0.362, and accuracy from 0.610 to 0.616. These trends indicate that incorporating longitudinal clinical context consistently helps—most notably at the one-year horizon where recent history is most informative—while at three years the gains are modest and trade off slightly against F1.

3.2 Limitations

This work has several limitations. Temporal information was incorporated through narrative prompts rather than explicit modeling, which may limit the ability to capture exact intervals or continuous disease trajectories. The reliance on manually designed and LLM-augmented templates also raises the possibility of prompt sensitivity. In addition, the dataset was derived from a single country screening programme, which may restrict generalizability to other populations, devices, or grading standards. Labels of progression were based on program follow-up and may contain noise from delayed attendance or unobserved care. Finally, while the model achieved consistent gains over baselines, absolute performance remained modest; larger and more diverse datasets, along with further external validation, calibration, and prospective evaluation, are required before clinical deployment.

4 Conclusion

We introduced a time-series vision–language model that integrates fundus images with structured clinical prompts to forecast diabetic retinopathy progression at 1-, 2-, and 3-year horizons. Conditioning on longitudinal history yielded consistent gains over image-only and tabular baselines, with best performance at shorter horizons. Beyond accuracy, the framework provides interpretability through clinical narratives and extensibility to richer temporal contexts.

References

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. doi: 10.1001/jama.2016.17216.

Daniel Shu Wei Ting, Charumathi Cheung, Gabriel W. 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. doi: 10.1001/jama.2017.18152.

Michael D. Abràmoff, Yinen Lou, Arezoo 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. doi: 10.1038/s41746-018-0040-6.

World Health Organization, Regional Office for Europe. Diabetic retinopathy screening: A short guide. increase effectiveness, maximize benefits and minimize harm, 2020. URL `https://iris.who.int/bitstream/handle/10665/336660/9789289055321-eng.pdf`. Accessed 4 Aug 2025.

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. doi: 10.2337/dc24-S012.

Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2:158–164, 2018. doi: 10.1038/s41551-018-0195-0.

Apeksha A. Bora, David S. W. Ting, Pearse A. Keane, and et al. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health*, 3(7):e476–e485, 2021. doi: 10.1016/S2589-7500(21)00059-0.

David Rom, Eyal Dok, Adiel Barak, and et al. Predicting the future development of diabetic retinopathy using a machine learning approach. *BMJ Open Ophthalmology*, 7(1):e001028, 2022. doi: 10.1136/bmjophth-2022-001028.

Liang Dai, Xinxing Xu, Rui Liu, et al. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*, 30:584–594, 2024. doi: 10.1038/s41591-023-02702-z.

Paul Nderitu, Zhen Qiu, Miguel Abós, and et al. Predicting 1-, 2- and 3-year emergent referable diabetic retinopathy and maculopathy from fundus photographs using deep learning systems. *Communications Medicine*, 4(1):68, 2024. doi: 10.1038/s43856-024-00611-1.

Yao Zhang, Xiaosong Wang, Ziyue Xu, et al. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of Machine Learning for Health (ML4H)*, 2020.

Tianshi Huang, Jianbo Jiao, Yixiao Zhang, et al. Gloria: A multimodal global-local representation learning framework for chest x-ray. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Owen Boecking et al. Biovil: Self-supervised vision-language pretraining for biomedicine. *arXiv preprint arXiv:2204.01461*, 2022.

Zhongzhu Wang et al. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Zhenyi Zhang et al. Biomedclip: A vision-language foundation model for biomedicine. *arXiv preprint arXiv:2306.07926*, 2023.

Edward Tiu, Eric Talius, Pooja Patel, and et al. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. doi: 10.1038/s41551-022-00936-9.

Yukun Zhou, Kai Yu, Jingke He, and et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 620:—, 2023. doi: 10.1038/s41586-023-06555-x.

Julio Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. ISSN 1361-8415.

Zhengyan Wang and et al. Retizero: Common and rare fundus diseases identification using a vision-language foundation model. *Nature Communications*, 2025. In press; preprint arXiv:2406.09317.

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