Probing Temporal Separability in Fluorescein Angiography Embeddings: A Foundation-Model Benchmark

Tymo van Rijn
Idiap Research Institute

May 11, 2026

Abstract

Fluorescein angiography (FA) is inherently dynamic. We investigate whether the latent representations of vision foundation models capture this temporal dimension. By evaluating six backbones through unsupervised Fisher probes and supervised GRU classification, we characterize how temporal binning strategies impact feature separability and downstream diagnostic accuracy. Our findings suggest that while embeddings encode significant temporal structure, finer-grained binning often yields diminishing returns for disease classification.

1 Introduction

Fluorescein angiography (FA) sequences are defined by the transit of dye through retinal vessels. While foundation models are often evaluated on static diagnostic tasks, their utility for longitudinal or sequence-based ophthalmic data remains under-explored. We propose a benchmark to quantify the "temporal awareness" of these models and determine if such signals translate to better clinical performance.

2 Methodology

2.1 Model Selection and Feature Extraction

We evaluate six vision backbones: DINOv2-Large, DINOv2-Small, MAE-Large, and three retina-specific fine-tuned models (RETFound-MAE, RETFound-DINOv2, and RETFound-Green). For each exam, frames are processed into a sequence of B embeddings using quantile and uniform binning strategies.

2.2 Temporal Probing and Classification

We utilize the Fisher discriminant ratio to assess how well embeddings from different temporal bins are clustered and separated in the feature space. Parallel to this, we train a single-layer GRU classifier on the binned sequences to predict five hyperfluorescence categories (None, Leakage, Staining, Pooling, Window Defect), allowing us to compare unsupervised separability with supervised utility.

3 Results

3.1 Multi-Bin Temporal Separability

We evaluate the separability of temporal phases across different binning resolutions (Table 1). We report both Raw Fisher scores (computed in the high-dimensional embedding space) and Fisher2D (computed in the 2D LDA projection space).

Table 1: Quantile binning metrics across models. Raw Fisher denotes full-space separability; F2D denotes separability in the 2D LDA projection.

	Q4		Q5		Q6	
Model	Raw	F2D	Raw	F2D	Raw	F2D
RETFound-MAE	0.1219	1.126	0.1217	1.000	0.1272	1.018
RETFound-DINOv2	0.0632	0.599	0.0645	0.549	0.0672	0.556
MAE-Large	0.0515	1.984	0.0505	1.802	0.0520	1.908
RETFound-Green	0.0345	1.083	0.0337	1.025	0.0349	1.095
DINOv2-Small	0.0297	1.106	0.0292	1.056	0.0300	1.109
DINOv2-Large	0.0260	2.062	0.0259	1.861	0.0271	1.964

3.2 Downstream Performance Metrics

Table 2 details the performance of the GRU head for the DINOv2-Large backbone. Interestingly, increasing the number of bins (B) does not lead to a linear increase in AUC-ROC or AP, suggesting a saturation point in temporal feature utility.

Table 2: Downstream GRU test metrics (mean ± std over multiple seeds).

Config	AUC-ROC	AP	F1
Quantile $B = 4$	0.775 ± 0.007	0.490 ± 0.003	0.453 ± 0.009
Quantile $B = 5$	0.763 ± 0.008	0.462 ± 0.009	0.434 ± 0.005
Uniform $B = 8$	0.762 ± 0.007	0.458 ± 0.010	0.447 ± 0.022

4 Discussion

The results highlight a divergence between geometric separability (Fisher scores) and task-specific utility. While RETFound-MAE exhibits the highest Raw Fisher scores, DINOv2-Large maintains high Fisher2D scores, indicating a more compact cluster structure in reduced dimensions. The decline in downstream performance with higher B values suggests that overly granular temporal binning may introduce variance that hinders the GRU's ability to generalize across exams with varying frame rates.

5 Conclusion

This benchmark provides a framework for evaluating the temporal dynamics of vision foundation models in medical imaging. We demonstrate that while temporal signal is robustly present, more granular phase identification is not a direct proxy for better disease classification. Future work should focus on temporal attention mechanisms to better leverage these latent signals.

References