Temporal Separability Benchmark

Study Proposal

Tymo van Rijn

March 12, 2026

1 Motivation

Temporal sequence models are most useful when the embedding space encodes meaningful temporal structure. Choosing the right foundation model and temporal binning strategy is currently done by expensive trial-and-error. A lightweight diagnostic tool that measures *how well* embeddings separate temporal states before training could substantially reduce this cost and provide principled, explainable model selection.

2 Research Question

> Does temporal separability in an embedding space, measured via a linear discriminant probe, predict the downstream performance of a temporal sequence model trained on that same embedding space?

3 Hypotheses

H1 If linear temporal separability measured by an LDA probe is a valid proxy for temporally useful structure in an embedding space, then embedding/binning configurations with higher separability scores will achieve better downstream performance when the same temporal sequence model is trained on them.

H2 If foundation models differ in how well they encode temporal structure, then the separability scores produced by the LDA probe will differ systematically across foundation models under comparable binning settings.

H3 If temporal binning influences how strongly temporal structure is represented independently of the embedding model, then varying the binning strategy or bin count while holding the embedding model fixed will lead to measurable changes in separability scores.

4 Method Sketch

Step	What
Input	Frame embeddings + elapsed time metadata
Binning	Multiple temporal binning strategies × multiple bin counts
Aggregation	Mean-pool frame embeddings within each bin to obtain bin-level vectors
Probe	LDA probe with Fisher separability ratio as the primary metric
Downstream	Select top-K^* and bottom-K^* configurations; train the same LSTM on ordered bin-level sequences
Evaluation	Grouped cross-validation or grouped splits by exam/patient; downstream performance measured with the primary task metric (e.g., macro-F1)
Validation	Test whether probe-based ranking correlates with downstream performance

5 Experimental Design

The study will benchmark whether a simple linear probe of temporal separability can serve as a useful diagnostic for downstream temporal modeling. To do so, multiple embedding configurations will be compared under controlled conditions.

5.1 Independent Variables

The benchmark will vary three main factors:

- **Foundation model:** different pretrained visual embedding models will be used to extract frame-level embeddings.
- **Temporal binning strategy:** elapsed time will be discretised using multiple strategies (e.g., uniform binning, quantile-based binning, or other clinically motivated schemes if available).
- **Number of bins:** multiple bin counts will be evaluated to test how coarse or fine temporal discretisation affects separability.

Each unique combination of foundation model, binning strategy, and bin count defines one experimental configuration, with the within-bin aggregation rule fixed to mean pooling in the present study.

5.2 Input Data

For each fluorescein angiography exam, frame-level embeddings and their associated elapsed time metadata will be used as input. The benchmark will operate on precomputed embeddings and existing metadata only. No new data collection, annotation, or foundation-model training will be performed.

5.3 Temporal Sequence Construction

For each fluorescein angiography exam, let $x_t \in \mathbb{R}^d$ denote the embedding of frame t, and let τ_t denote its elapsed time. Given a temporal binning scheme with B bins, each frame is assigned to one bin based on τ_t.

For each bin $b \in \{1, \ldots, B\}$, let S_b be the set of frame embeddings assigned to that bin. A single bin-level representation z_b is then constructed by mean pooling:

$$z_b = \frac{1}{|S_b|} \sum_{x_t \in S_b} x_t.$$

This produces an ordered sequence of bin-level representations

$$(z_1, z_2, \ldots, z_B),$$

which is used as the input sequence for the downstream LSTM.

In this way, temporal bins are not fed into the LSTM as labels or as randomly selected frames. Instead, each bin is converted into a single aggregated embedding vector, so that the temporal discretisation used in the probe is also reflected in the downstream sequence model.

If a bin is empty for a given exam, the corresponding time step will be handled consistently through padding and masking in the downstream model.

5.4 Probe Stage

For every configuration, frame-level embeddings are first assigned to temporal bins according to the selected binning strategy and number of bins, after which bin-level representations are constructed through within-bin aggregation.

The probe dataset therefore consists of aggregated bin-level vectors pooled across exams, with the corresponding temporal bin index used as the class label. A linear discriminant analysis (LDA) probe is then applied to assess how well these temporal states can be separated in the representation space.

The primary probe metric will be the Fisher separability ratio, used to rank configurations from high to low temporal separability.

5.5 Downstream Stage

After ranking all configurations based on probe separability, a subset of top-K and bottom-K configurations will be selected for downstream validation. For each selected configuration, the ordered sequence of aggregated bin-level embeddings will be used as input to the same downstream LSTM for the target temporal modeling task.

This design allows us to test whether strong linear temporal separability is associated with better downstream temporal-model performance, while avoiding the computational cost of training a temporal model for every possible configuration.

5.6 Controlled Factors

To ensure fair comparison across configurations, the following factors will be kept constant wherever possible:

- the downstream temporal model architecture,
- the train/validation/test splitting protocol,
- the evaluation metric,
- the training procedure and optimisation settings,
- the preprocessing pipeline applied before embedding extraction or downstream modeling.

5.7 Validation Protocol

Evaluation will be performed using grouped cross-validation or grouped data splits at the exam and/or patient level, so that frames from the same exam or patient do not leak across training and evaluation sets. This is necessary to obtain a realistic estimate of generalisation performance.

6 Analysis Plan

The analysis will evaluate whether temporal separability, measured by a linear probe, is informative for downstream temporal modeling performance. All reported results will be computed under the grouped validation protocol described above, and scores will be aggregated across folds where applicable.

6.1 Primary Quantities

For each experimental configuration, two main quantities will be obtained:

- **Probe score:** the Fisher separability ratio produced by the LDA probe, measuring how well temporally binned representations are linearly separable.
- **Downstream score:** the performance of the downstream LSTM on the target task, measured using the primary evaluation metric (e.g., macro-F1).

  Thus, each configuration c yields a pair

$$(s_c, p_c),$$

where s_c denotes the probe separability score and p_c denotes downstream performance.

6.2 Analysis for H1

H1 states that configurations with higher linear temporal separability should also yield better downstream temporal-model performance.

To test this hypothesis, all evaluated configurations will first be ranked according to their probe scores s_c. For the subset of configurations selected for downstream validation, the relationship between probe score and downstream score will then be assessed.

The primary analysis for H1 will be a rank-based correlation between s_c and p_c, using Spearman's ρ. A positive correlation would indicate that configurations with stronger temporal separability tend to perform better downstream.

In addition, a top-versus-bottom comparison will be performed by comparing the downstream scores of the highest-ranked and lowest-ranked configurations according to the probe. If the top-ranked configurations consistently outperform the bottom-ranked configurations, this will provide further support for H1.

6.3 Analysis for H2

H2 states that foundation models differ in how well they encode temporal structure.

To test this hypothesis, probe scores will be compared across foundation models under matched binning settings. For each foundation model, the distribution or average of separability scores across comparable binning configurations will be examined.

Support for H2 would be provided if some foundation models consistently yield higher separability scores than others under the same temporal discretisation settings. Where appropriate, these differences will also be inspected in relation to downstream performance to determine whether better separability at the probe level is accompanied by better temporal-model utility.

6.4 Analysis for H3

H3 states that temporal binning influences how strongly temporal structure is represented independently of the embedding model.

To test this hypothesis, probe scores will be compared across binning strategies and bin counts while holding the foundation model fixed. This analysis will examine whether separability changes systematically as a function of temporal discretisation.

Support for H3 would be provided if varying the binning strategy or the number of bins leads to consistent and measurable changes in probe separability within the same foundation model.

6.5 Interpretation Criteria

The hypotheses will be interpreted as follows:

- **H1 is supported** if higher probe separability is associated with better downstream performance, for example through a positive Spearman correlation and/or clear top-versus-bottom performance differences.
- **H2 is supported** if probe separability differs systematically across foundation models under comparable binning conditions.
- **H3 is supported** if probe separability changes systematically when the binning strategy or bin count is varied while the foundation model is kept fixed.

6.6 Negative and Mixed Outcomes

Several outcome patterns are possible. If separability differs across configurations but does not align with downstream performance, then the probe may still reveal structural differences in the representation space without being a reliable predictor of downstream usefulness. If no meaningful differences are observed at all, this would suggest that linear temporal separability is not a strong diagnostic signal in this setting. Such negative or mixed findings would still be informative, as they would clarify the limitations of the proposed benchmark.

7 Expected Outcome

We expect that configurations with stronger linear temporal separability will also achieve better downstream performance in temporal modeling. If this is confirmed, the separability probe may serve as a cheap and general-purpose pre-selection tool for embedding/binning configurations. If this is not confirmed, the result will still be informative, as it would clarify the limitations of linear temporal probing as a proxy for downstream usefulness.

8 Scope — What We Are NOT Doing

- No new foundation model training or fine-tuning.
- No non-linear probes (no kernel methods, no neural probes).
- No clinical outcome prediction.
- No dataset collection or annotation.
- No deployment or real-time inference.

*$top\text{-}K$: The best-performing configurations based on the probe metric.
*$bottom\text{-}K$: The worst-performing configurations based on the probe metric.