Analysis Report on Temporal Encoding Techniques

Tymo van Rijn

March 2026

1 Motivation

After a very informative meeting with Oscar this Tuesday (17 March), he pointed out something quite important; *Tunnel Vision*. As of right now, I am mainly looking into binning, which is not necessarily a bad thing, but I should also be open for other temporal encoding solutions. Which I think was not really happening anymore, I was so focused on one particular solution that I created a tunnel on this solution.

I am glad Oscar pointed this out to me, because this made me realise I should keep up with the recent literature about encoding temporal information more. Which is exactly what I am writing this document for, it is meant to analyse recent papers on encoding temporal information, to make sure I am not trying to reinvent the wheel with the ongoing study.

2 Studies

2.1 L-MAE: Longitudinal masked auto-encoder with time and severity-aware encoding for diabetic retinopathy progresiion prediction (*2025, Zeghlache et al.*)

This is one of the closest paper to our case as far as I am aware. The core idea is to extend MAE-style self-supervised pretraining with a **time-aware position embedding** that uses actual intervals between visits, plus a **disease progression-aware masking strategy** so the model learns changes that matter clinically rather than only generic image context. For **FA**, this is highly relevant because our sequences are not just ordered frames; they reflect a clinical progression process with meaningful time gaps and stage-specific pathology.

2.2 Harnessing the power of longitudinal medical imaging for eye disease prognosis using Transformer-based sequence modeling(*2024, Holste et al.*)

This paper proposed a **Longitudinal Transformer for Survival Analysis (LSTA)** on sequences of fundus photographs collected over **long, irregular time periods.** The direct relevance is that it shows transformers sequence modelling can work in ophthalmic longitudinal imaging when visit spacing is irregular and the signal is progression-oriented. Even though FA is a different modality and much shorter-horizon than follow-up fundus exams, this paper is a good medical proof point that ophthalmic temporal structure is worth modeling explicitly.

2.3 T-Rep: Representation Learning for Time Series Using Time-Embeddings (*2024, Fraikin et al.*)

T-Rep explicitly learns **vector embeddings of time** alongside the feature extractor and uses them in self-supervised pretext tasks to capture trend, periodicity, and distribution shifts. Conceptually, this is one of the clearest papers on "temporal encoding" as a first-class representation problem. For the FA, this transferable idea is simple and strong: do not encode only the frame content; encode the timestamp or elapsed time as an ambedding that interacts with the visual features.

2.4 LOMIA-T: A Transformer-based LOngitudinal Medical Image Analysis framework for predicting treatment response of esophageal cancer (*2024, Sun et al.*)

LOMIA-T learns from **pre/post longitudinal image pairs** using a treatment-response contrastive loss and then fuses the latent representations via **cross-attentions**. This is a strong example of temporal encoding through **contrastive evolution modeling** rather than only plain sequence processing. For FA, that suggests a useful direction: contrast representations of early vs midle vs late phases, or adjacent temporal bins, to force the encoder to learn "how the image evolves", not just what it looks like at one timepoint.

2.5 XTSFormer: Cross-Temporal-Scale Transformer for Irregular-Time Event Prediction in Clinical Applications (*2025, Xiao et al.*)

XTSFormer introduces a **Feature-based Cycle-aware Time Positional Encoding (FCPE)** and a **hierachical multi-scale temporal attention** mechnaism for irregularly timed data. Although the domain is clinical event sequences rather than images, the technical idea is very relevant: positional encoding should not be naive when timing is irregular and structure exists at multiple temporal scales. For FA, this paper is useful to design a more principled temporal encoder than standrd position embeddings.

3 Conclusion

This short analysis shows that temporal encoding is a much broader topic than only binning. Across the reviewed studies, the main shared idea is that temporal information should usually be modeled **explicitly**, rather than being treated as a simple ordering of frames. Different papers do this in different ways: by adding learned time embeddings, by using time-aware positional encodings, by modeling change between timepoints, or by learning temporal structure at multiple scales.

For our fluorescein angiography (FA) problem, this is highly relevant. An FA examination is not just a collection of retinal images. It is a temporally evolving process in which the moment of acquisition matters. Early, middle, and late phases can contain different information, and certain pathological patterns are only meaningful in relation to how they appear or change over time. This means that a model that only looks at image content may miss part of the signal, while a model that also receives temporal information may be able to learn a more informative representation.

Based on the reviewed papers, the most practical implementation for our case would be to start with a relatively simple but principled temporal encoding pipeline. First, each FA frame or temporal bin can be converted into a feature vector using an image encoder. Second, for each frame or bin, a time representation can be added, for example based on elapsed time since the start of the examination or since dye injection. Third, the sequence of image features and time

embeddings can be processed by a temporal model such as a Transformer. Finally, the resulting temporal representation can be used for the downstream task, such as predicting the exam-level label or classifying temporal pathology patterns.

In this context, the literature suggests several concrete directions. A first option is a **time-embedding approach**, where the timestamp is learned as an embedding and combined with the visual features. A second option is a **phase-comparison approach**, where the model explicitly learns how early, middle, and late phases differ from each other. A third option is a **multi-scale temporal approach**, where the model captures both short-term and longer-term temporal structure. Among these, the time-embedding approach appears to be the most feasible and interpretable starting point for our current study, because it can be integrated into the existing pipeline without requiring a fully new framework.

Therefore, the main conclusion of this analysis is that binning should not be viewed as the only reasonable solution. It remains a useful baseline, but recent literature indicates that temporal information can also be encoded more directly and potentially more effectively. For the ongoing study, this suggests that the most sensible next step is not to abandon binning entirely, but to compare it against at least one explicit temporal encoding strategy. In that way, the final design can be based on evidence rather than on tunnel vision toward a single method.