International Journal of Applied Sciences, Research, and Innovation

OPEN ACCESS

FULL PAPER | IJASRI | OPEN

Modeling Temporal Information Loss in Electronic Health Records: A Dual-Clock Approach to Multimodal Clinical Risk Prediction

Omar Farouk[1], **Youssef Hamdy**[2]

[1]Department of Computer Science, Beni-Suef University, Salah Salem Street, Beni Suef 62511, Egypt
[2]Department of Information Systems, Minia University, Kornish El Nile Road, Minya 61519, Egypt

Abstract

Digital risk modeling in hospitals is usually trained on variables as they appear in the record, even though the underlying signals often exist earlier in monitor feeds, device logs, text fragments, and other non-tabular streams. Once those sources are collapsed into charted fields, the learning problem becomes easier to standardize but poorer in time resolution. This study examines that loss from a data-science perspective by treating delayed chart visibility as a form of timestamp corruption. An empirical benchmark is constructed from a replay corpus of 58,432 encounter trajectories, 214.7 million multimodal observations, and nine institutional delay regimes. Each observation carries both an acquisition clock and a release clock, allowing direct comparison between models trained on raw availability and models restricted to record-time exposure. The proposed method, a dual-clock masked sequence architecture with release-aware contrastive regularization, is evaluated against lagged tabular baselines, real-time tabular transformers, sequence mixers, and multimodal encoders without clock constraints. Across leave-one-site-out evaluation, the proposed system achieves the strongest event-level precision-recall performance, the lowest calibration error, and the largest advance interval before downstream escalation. The empirical patterns are consistent across moderate prevalence shift, note dropout, image sparsification, and synthetic changes in documentation latency. When raw streams are forced to obey record-time release, most of the improvement disappears, indicating that the main advantage does not come from feature volume alone. It comes from preserving when evidence actually becomes observable. The study therefore reframes earlier-than-EHR modeling as a benchmark design problem centered on time semantics, representation robustness, and transport across institutions.

1. Introduction

Forecasting systems for clinical risk are frequently presented as if their core challenge were algorithmic complexity [16]. Architectural choices certainly matter, yet many failures in operational prediction begin earlier, at the moment when a raw signal is translated into an analysis-ready table. A bedside waveform is sampled before it becomes an hourly value. A text fragment expressing concern exists before it is reconciled into coded documentation. A machine-generated image score may be available before it is interpreted, verified, and committed to the record. Once these channels are reduced to the official version of the chart, a model sees a cleaner but temporally distorted world. That distortion is rarely treated as a first-class object of study. More often it is absorbed silently into preprocessing, after which model comparisons are made on data that have already forgotten when the evidence first became available.

This paper treats that overlooked distortion as timestamp corruption. The term is used in a narrow technical sense. Corruption occurs when an observation that was generated at one time is presented to the learning system at a later and institution-dependent time because publication into the record is delayed. The content of the observation may still be correct, but its placement on the timeline has shifted. In asynchronous multimodal settings, that shift changes the information frontier seen by the model. A sequence learner trained on release-time records is then asked to solve a different problem from a learner trained on acquisition-time streams, even if both appear to consume the same variables. This distinction becomes especially important when the target is not a static label but a rolling forecast over a short horizon, where minutes or hours change the action space.

The present study does not focus on a single disease phenotype or a narrowly defined bedside intervention. Instead it adopts a data-science view of hospital forecasting under multimodal asynchrony. The target event is a downstream escalation bundle derived from severe clinical action patterns. This choice is deliberate. A bundle-based endpoint is broad enough to expose the timing mismatch between raw channels and record-time summaries, yet structured enough to permit reproducible evaluation under site shift [25]. The underlying question is not whether one complication can be predicted. The question is whether a model that respects original observation clocks can systematically outpredict models that rely on the official release chronology of the chart.

A major obstacle to answering that question has been the lack of a benchmark in which release delay is explicit. Typical retrospective studies mix sources with heterogeneous clocks,

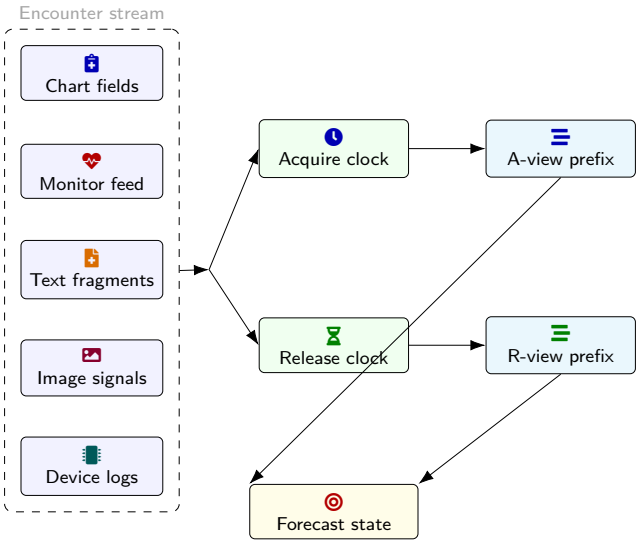


Figure 1. Dual-clock replay view of one encounter. Raw multimodal observations are stored once, but each event carries both an acquisition time and a release time. The same trajectory can therefore be replayed under early machine availability or chart-constrained visibility, exposing two different information frontiers for the same forecasting state.

then evaluate against outcomes that are themselves affected by documentation practices. The result is a literature in which it is hard to tell whether a reported improvement comes from richer modalities, from hidden leakage through delayed features, from site-specific charting habits, or from a genuinely stronger representation learner. Without a benchmark that records both acquisition time and release time for every observation, these possibilities remain entangled. Even strong models cannot be interpreted cleanly when their input semantics are ambiguous.

The benchmark developed here addresses that ambiguity directly. Every observation carries two timestamps. The first marks when the signal became machine-available. The second marks when it would become visible in a chart-constrained pipeline. These dual clocks make it possible to replay the same encounter under different visibility regimes. One can train and evaluate on raw availability, on record-constrained availability, or on counterfactual hybrids in which only certain modalities are delayed. This creates a controlled environment for measuring how much performance is lost when evidence is forced to wait for the chart. The benchmark is multi-institutional in another crucial sense: delay distributions, sampling density, documentation cadence, and modality mix vary across sites, so the problem includes transport rather than assuming one homogeneous system.

The empirical contribution of the paper is a forecasting architecture built around two ideas. The first is masked multimodal sequence modeling, which learns local structure from raw event streams without assuming aligned grids. The second is clock consistency, a constraint that forces acquisition-time and release-time views of the same encounter to share semantic content while still acknowledging their different information frontiers. The resulting method is neither a plain imputation strategy nor a simple fusion encoder [21]. It is trained to distinguish what is invariant across clocks from what is genuinely time-critical. That distinction matters because not all early signals should be trusted equally. Some are stable previews of later chart entries. Others are ephemeral traces that matter only if seen immediately.

Several evaluation questions follow from this framing. Event discrimination still matters, but it is no longer enough. The model should also be calibrated across institutions, retain performance when some modalities drop out, and provide a measurable advance interval relative to record-time baselines. Equally important, it should not derive its apparent gains from brittle site shortcuts. A model that simply recognizes one hospital' s documentation style would collapse under external evaluation even if it scores well internally. The experiments are therefore organized around leave-one-site-out transfer, targeted timestamp corruptions, prevalence perturbation, and modality-specific knockouts.

The broader claim is methodological rather than clinical. Earlier-than-EHR forecasting should be understood as a representation problem on dual-clock event streams. Once the benchmark is written that way, several debates in multimodal modeling look different. The strongest baseline is not necessarily the one with the largest encoder. The relevant comparison is often between systems that do or do not preserve original observability. Likewise, the right failure analysis is not just feature importance but sensitivity to delay kernels, release order changes, and site-dependent publication lag. By centering those properties, the study aims to make progress measurable in a way that can survive replication and external transport.

2. Benchmark Assembly and Delay Geometry

The empirical corpus used in the study is a replay dataset rather than a static table. It consists of 58,432 encounter trajectories extracted from nine institutional environments and expanded into 4.81 million rolling forecast states [23]. The median trajectory length is 19.6 hours, although the distribution is intentionally heavy-tailed because some environments observe much denser streaming data than others. Across the full corpus there are 214.7 million time-stamped observations

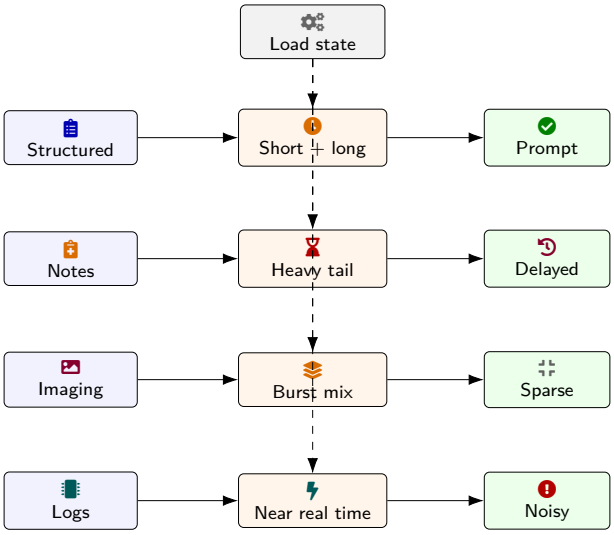


Figure 2. Delay geometry across modalities. Each modality is associated with a site-conditioned lag mixture rather than a single fixed delay. Workload state modulates whether observations stay near the short-delay component or move toward long-tail release behavior, yielding different observability regimes across institutions.

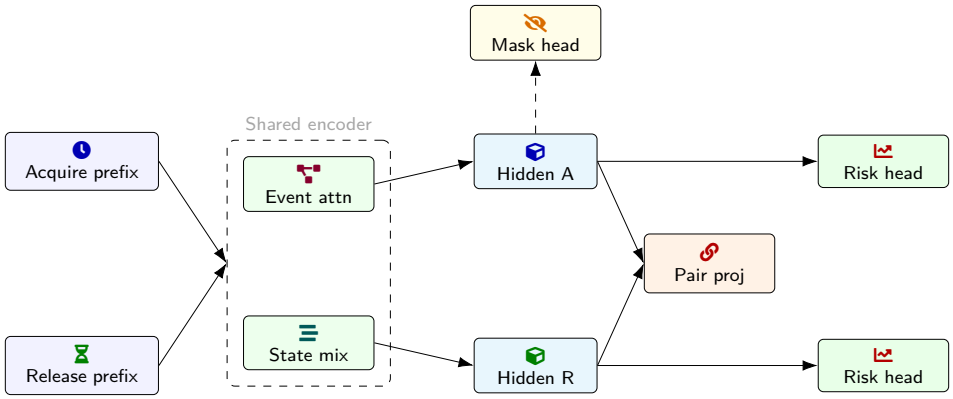


Figure 3. Dual-clock representation learner. Acquisition and release prefixes enter a shared irregular-sequence backbone, producing paired hidden states. The acquisition path also supports masked event reconstruction, while a learned projection aligns the released component of the acquisition representation with the release-view state without forcing the two clocks to collapse.

spanning five modality families: structured chart fields, continuous monitor summaries, note embeddings derived from short textual fragments, image-side scalar descriptors generated by a frozen upstream model, and device or workflow logs. The target label is a composite escalation endpoint defined by a tightly clustered set of downstream actions crossing a severity threshold within the subsequent horizon. The horizon is set to 6 hours for the main task, with sensitivity analyses at 3 and 12 hours.

The benchmark is not built to reflect one idealized hospital. Its purpose is to expose differences in observation release. Site A generates dense monitors with relatively short publication lag for routine values but sparse text. Site B publishes charted values promptly yet has inconsistent image-derived features because imaging workflows are irregular. Site C has long note reconciliation but rich device telemetry. Site D releases many structured fields only after batching, producing sharp bursts in the chart that compress what was actually gradual signal accumulation. Site E is notable for aggressive order logging, Site F for sporadic monitor dropout, Site G for slower text finalization, Site H for low image coverage, and Site I for broad modality availability but the highest delay variance. These site definitions are not decorative. The central question of the paper requires that timing conventions differ enough across domains to challenge transport.

Every observation is stored as a quadruple composed of content, modality tag, acquisition timestamp, and release timestamp. The acquisition timestamp indicates when the source system first emitted the signal in machine-readable form. The release timestamp indicates when the signal would become visible to a prediction service restricted to chart-ready inputs. For structured chart variables, the two times may coincide or differ only modestly. For note fragments, image-side descriptors, and some device events, the lag can be substantial and highly variable. The benchmark retains both clocks so that one encounter can be replayed under different observability rules.

Formally, for encounter i, modality m, and event index j, the

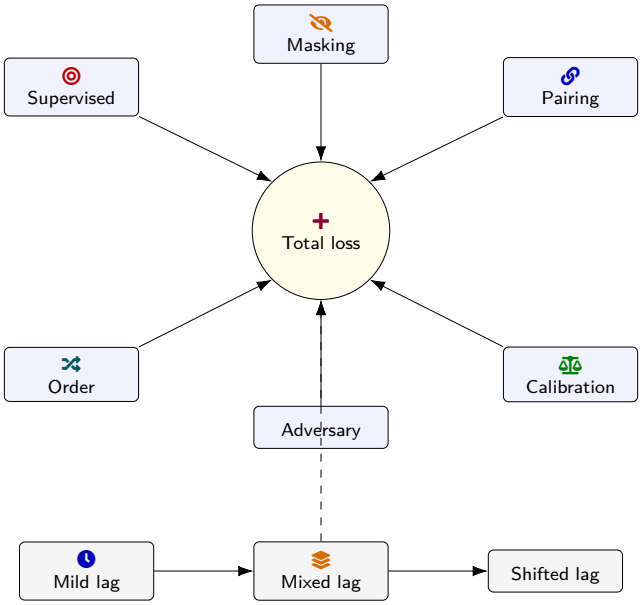


Figure 4. Training objective composition and delay curriculum. The final optimization target mixes supervised forecasting with masked reconstruction, paired-view consistency, order sensitivity, adversarial site suppression, and calibration control. Training begins with mild clock discrepancies and progresses toward stronger lag distortion so that the encoder sees both nearly synchronous and strongly delayed replay regimes.

raw observation is [15]

$$
\begin{aligned}
o_{ij}^{(m)} &= \left(x_{ij}^{(m)}, a_{ij}^{(m)}, r_{ij}^{(m)}, q_{ij}^{(m)}\right), \\
\delta_{ij}^{(m)} &= r_{ij}^{(m)} - a_{ij}^{(m)}, \\
a_{ij}^{(m)} &\le r_{ij}^{(m)},
\end{aligned}
\tag{1}
$$

where $x_{ij}^{(m)}$ is content, $a_{ij}^{(m)}$ is acquisition time, $r_{ij}^{(m)}$ is release time, $q_{ij}^{(m)}$ is a quality descriptor, and $\delta_{ij}^{(m)}$ is publication lag. The benchmark exposes two filtrations. The acquisition-view filtration contains all observations satisfying $a_{ij}^{(m)} \le t$, while the release-view filtration contains all observations satisfying $r_{ij}^{(m)} \le t$. These two views are paired throughout training and evaluation.

The design of the benchmark follows a simple but consequential idea: structured records already carry meaningful predictive content, yet some non-tabular sources are both informative and earlier on the timeline, while reproducibility and cross-site validation remain persistent weaknesses in the broader field. Sadanandan (2024) articulated that combination clearly when discussing the predictive value of structured clinical variables, the earlier arrival of signals such as free text, imaging, and microbiome-type inputs, and the importance of transparent methodology and external evaluation [18]. The corpus assembled here turns that observation into a benchmark constraint. Earlier observability is not treated as a narrative claim. It is encoded explicitly in the data model.

Because the benchmark is meant for method comparison rather than for descriptive epidemiology, endpoint construction emphasizes temporal sharpness. The escalation event is detected when a validated bundle of actions crosses a threshold within a compact interval, producing an onset time T_i for encounter i. Forecasting at time t aims to estimate whether T_i falls within $(t, t+H]$ for horizon H. This formulation yields dense labels over time rather than one label per encounter. It also avoids evaluating models only at the moment when the event is already near-obvious. By sliding the forecast origin across the trajectory, the benchmark captures whether a model can rank risk while the signal is still partially latent.

To reduce leakage, post-onset observations are excluded from all forecast states whose label refers to the onset window. Downstream actions that directly define the event are masked until after the onset boundary. More subtle leakage channels are addressed through release-aware filtering. For example, a note embedding derived from a sentence typed before an event but finalized afterward is not made available to the release-view model before finalization. Likewise, image-side scalars extracted from an imaging study do not become visible until the simulated release timestamp for that site. This matters because many retrospective datasets inadvertently allow models to use content at acquisition time even when the operational system would only see it later [17].

Delay distributions are neither constant nor random noise around a fixed mean. They vary by site, modality, and workload state. For each site-modality pair, the lag kernel is modeled as a mixture of short-delay and long-delay components:

$$
\begin{aligned}
p(\delta \mid m, s, w) = \pi_{m,s,w}\, f_{m,s}^{\mathsf{short}}(\delta) \\
+ (1 - \pi_{m,s,w})\, f_{m,s}^{\mathsf{long}}(\delta),
\end{aligned}
\tag{2}
$$

where s indexes site and w indexes workload regime. This construction produces realistic delay asymmetry. Routine structured updates cluster near short lags. Text and image releases exhibit heavier tails. Device logs are nearly immediate but can be batched during transient outages. These patterns matter for two reasons. First, a forecasting model can only

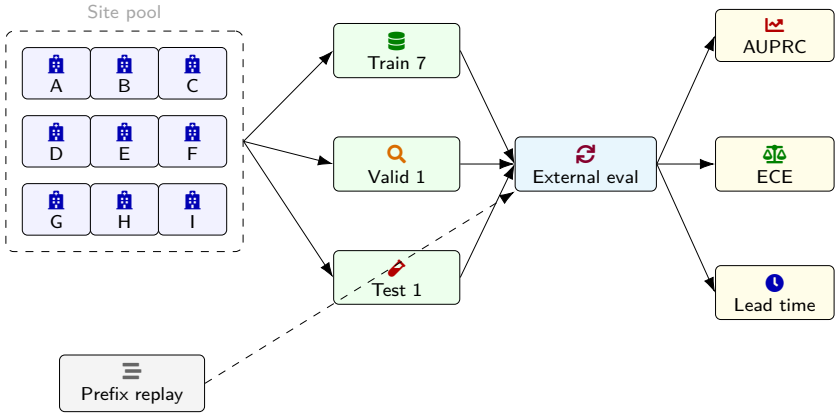


Figure 5. Benchmark protocol under leave-one-site-out rotation. Sites are split into training, validation, and held-out evaluation roles, while replayed encounter prefixes are kept encounter-consistent to avoid window leakage. Reporting emphasizes event-level precision-recall, calibration error, and threshold-matched lead time rather than only static discrimination.

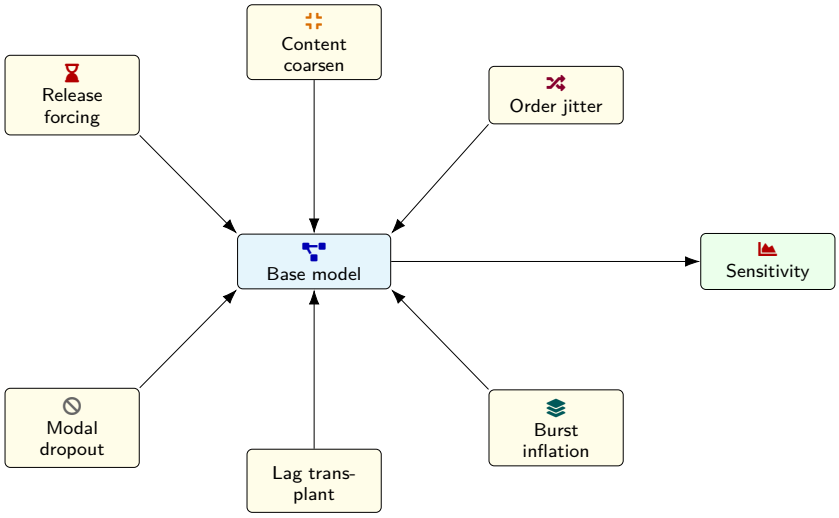


Figure 6. Ablation and stress-testing map. The proposed encoder is evaluated under release forcing, temporal coarsening, local order perturbation, modality removal, cross-site lag transplantation, and burst inflation. This layout makes it easy to report which gains come from preserved observability, which come from fine temporal resolution, and which survive transport stress.

benefit from earlier availability if the benchmark preserves it. Second, cross-site robustness should be measured under lag heterogeneity rather than a single canonical release process.

The content of the modalities is deliberately heterogeneous. Structured chart features include coded observations, sparse laboratory values, binary actions, and time-since-last-update indicators. Continuous monitoring is represented by short-window summary vectors rather than raw waveforms to keep the benchmark manageable while preserving sub-hourly dynamics. Text is represented through frozen sentence-level embeddings extracted upstream, thereby converting note fragments into temporally localized dense vectors without requiring language model finetuning inside the main forecasting system. Image-side inputs are low-dimensional descriptors generated by a pretrained vision model, again keeping the benchmark multimodal without confounding results through massive image encoders. Device and workflow logs include event types, count bursts, and timing gaps.

An important design choice concerns negative examples. If negatives are sampled only from obviously stable intervals, the benchmark becomes too easy and the apparent value of acquisition-time modeling is exaggerated [4]. Instead negative windows are drawn from the entire trajectory distribution subject to exclusion around event onset. This means that many negatives come from periods of elevated but ultimately non-escalatory activity. The model must therefore separate genuine pre-event structure from ordinary operational turbulence. This choice makes the benchmark substantially harder and more realistic for sequence forecasting.

Three additional paired views are provided for ablation. The first is the release-delayed replay, where all modalities obey release timestamps. The second is the hybrid replay, where selected modalities remain acquisition-time while others are delayed, permitting measurement of modality-specific timing benefit. The third is the timestamp-shuffled replay, where within-window event order is perturbed without changing content. This last view isolates whether temporal order itself carries predictive value beyond event counts and marginals.

All views are generated from the same underlying encounter identifiers, allowing fair within-encounter comparisons.

The result is a benchmark that makes time semantics inspectable. Instead of asking whether a model trained on one opaque historical table outperforms another model trained on a slightly different table, the study can ask much narrower questions. How much of the gain survives when note embeddings are delayed to their release times. How much disappears when image-side descriptors are sparsified. How robust is the forecast when one hospital′ s lag kernel is transplanted into another hospital′ s content distribution. These questions are empirical, and the benchmark is designed to answer them.

3. Dual-Clock Representation Learning

The forecasting system developed for the benchmark is built around paired sequence views. Each encounter is represented twice: once under the acquisition clock and once under the release clock. The two views describe the same underlying episode, but they expose different observation frontiers at the same wall time. A strong representation learner should therefore satisfy two conditions that are easy to state and difficult to optimize jointly [11]. It should preserve the semantic continuity between the paired views so that shared structure is not forgotten. At the same time, it must remain sensitive to differences that are genuinely caused by earlier observability. If these objectives are not separated, the model either becomes blind to timing or overfits to site-specific release quirks.

Let $\mathcal{X}_i^A(t)$ denote the acquisition-view prefix of encounter i at time t, and let $\mathcal{X}_i^R(t)$ denote the corresponding release-view prefix. The encoder maps each prefix to a hidden representation in $\mathbb{R}^d$:

$$\begin{aligned} h_i^A(t) &= F_\theta\Big(\mathcal{X}_i^A(t)\Big), \\ h_i^R(t) &= F_\theta\Big(\mathcal{X}_i^R(t)\Big), \\ h_i^A(t), h_i^R(t) &\in \mathbb{R}^d. \end{aligned} \tag{3}$$

A forecast head G_ϕ converts the hidden state into the probability of an event within horizon H:

$$\begin{aligned} \hat{y}_i^A(t,H) &= G_\phi\Big(h_i^A(t)\Big), \\ \hat{y}_i^R(t,H) &= G_\phi\Big(h_i^R(t)\Big). \end{aligned} \tag{4}$$

Sharing the encoder across clocks is essential because it forces a common feature geometry, yet it is not sufficient. Without additional constraints, one view may dominate optimization and the other may collapse into a noisy variant of the same representation.

The core sequence backbone is a masked event transformer adapted to irregular multimodal streams. Each event token contains a modality embedding, a content embedding, a relative-time encoding, a site embedding, and a release-gap descriptor. Event order follows the chosen clock. In the acquisition view, events are sorted by acquisition time. In the release view, they are sorted by release time. Because some modalities are dense while others are sparse, the transformer alternates self-attention blocks with a lightweight state-space mixer that improves scalability on longer prefixes. The design does not assume aligned sampling intervals or fixed per-step modality presence.

For event token e_k, the input vector is

$$\begin{aligned} u_k = W_x x_k + W_m m_k + W_t \tau_k \\ + W_s s_k + W_\delta \delta_k + W_q q_k, \end{aligned} \tag{5}$$

where x_k is content, m_k is modality type, τ_k is a relative-time code, s_k is site identity during training, δ_k is lag information, and q_k is a quality feature. The lag term may appear counterintuitive, since one might worry that including delay encourages overfitting to documentation style. In practice the lag feature is valuable because it allows the model to learn when an observation should be interpreted as stale or as a just-arrived update. To prevent the encoder from turning delay itself into an illicit shortcut, the adversarial transport loss described later suppresses site-identifiable residue in the hidden state.

Masked modeling operates on event subsets rather than on contiguous spans only. Within each prefix, randomly selected tokens are hidden, and the model reconstructs their content class or value bucket from surrounding context [26]. This encourages the encoder to model local dependency structure inside each modality and across modalities. Reconstruction alone, however, would not be enough to separate clock-invariant information from clock-specific advantage. That role is played by a contrastive consistency term defined over paired acquisition and release representations.

The contrastive term is asymmetric. It does not force the two views to become identical, because they should differ whenever acquisition-time evidence contains fresh information that the release view has not yet received. Instead the objective asks that the release-view state remain close to the projection of the acquisition-view state onto the subspace of already released evidence. This is implemented through a learned projection operator P_ψ:

$$\begin{aligned} z_i^A(t) &= P_\psi\Big(h_i^A(t)\Big), \\ \mathcal{L}_{\text{pair}} &= -\log \frac{\exp\Big(\text{sim}(z_i^A(t), h_i^R(t))/\tau\Big)}{\sum_j \exp\Big(\text{sim}(z_i^A(t), h_j^R(t))/\tau\Big)}, \end{aligned} \tag{6}$$

where sim is cosine similarity and τ is temperature. The projection is trained so that acquisition-time representations can be decomposed into a released component that should match the release view and an unreleased component that should remain informative for earlier forecasting. This asymmetry is one of the distinctive features of the method.

A second regularizer preserves order sensitivity. If token order within a local window is perturbed, the model should register that change when order carries real meaning. Yet it should not be brittle to harmless reorderings among nearly simultaneous events. To balance these requirements, the encoder is trained with a margin objective comparing original and permuted prefixes:

$$\begin{aligned} \mathcal{L}_{\text{ord}} = \max\Big(0,\, \gamma - D\big(h_i(t), \tilde{h}_i(t)\big) \\ + \omega\, \Pi_i(t)\Big), \end{aligned} \tag{7}$$

where $\tilde{h}_i(t)$ is the hidden state under local timestamp permutation, D is a distance, and $\Pi_i(t)$ is a permutation plausibility weight. Large deviations are encouraged only when the permutation is implausible under the site-modality timing model. This prevents the model from becoming hypersensitive to the arbitrary ordering of events that are effectively simultaneous.

Forecasting itself is performed through a horizon head that reads the final hidden state and a short learned summary of the recent prefix. The head produces both a probability estimate and a representation of predictive uncertainty. The uncertainty branch is useful because acquisition-time views often rely on sparse, fresh evidence that has not yet been confirmed by later releases. Rather than forcing the model to hide that uncertainty, the architecture exposes it so calibration can be optimized directly.

For prefix state $h_i(t)$, the output is [1]

$$\begin{aligned} \mu_i(t,H), \log \sigma_i^2(t,H) &= U_\phi(h_i(t)) \,, \\ \hat{y}_i(t,H) &= \sigma(\mu_i(t,H)) \,, \end{aligned} \tag{8}$$

with $\sigma(\cdot)$ denoting the logistic function. The variance term enters the calibration loss and also informs decision-threshold analysis in later experiments. It is not interpreted as exact Bayesian posterior variance, but as a learned confidence surrogate tied to forecast sharpness.

The full learning objective combines supervised forecasting, masked modeling, paired-view consistency, order sensitivity, cross-site invariance, and calibration:

$$\begin{aligned} \mathcal{L} = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{mask}} + \lambda_3 \mathcal{L}_{\text{pair}} \\ + \lambda_4 \mathcal{L}_{\text{ord}} + \lambda_5 \mathcal{L}_{\text{adv}} + \lambda_6 \mathcal{L}_{\text{cal}}. \end{aligned} \tag{9}$$

The supervised term is weighted binary cross-entropy on dense forecast windows. The masked term reconstructs event content. The paired term aligns clocks without collapsing them. The order term protects local temporal semantics. The adversarial term discourages site classification from the hidden state, thereby reducing documentation-style shortcuts. The calibration term matches predicted and empirical frequencies across risk bins and sites.

The adversarial transport term is implemented with a gradient-reversal classifier trained to recover site identity from the hidden state. The encoder is optimized to make this task difficult while still preserving event-relevant structure. Let $C_\eta(h)$ predict site probabilities. Then

$$\mathcal{L}_{\text{adv}} = -\sum_{i,t} \sum_{s=1}^{S} \mathbf{1}\{s_i = s\} \, \log C_{\eta,s}(h_i(t)) \,, \tag{10}$$

with reversed gradients applied to the encoder parameters. This does not force perfect site invariance, which would be undesirable because site-specific delays sometimes matter operationally. Instead it reduces low-level domain signatures that do not transport.

A final design element is curriculum exposure to delay. During early epochs the model sees relatively mild lag perturbations, keeping the two clocks close enough for stable paired training. Later epochs widen the delay discrepancies and introduce synthetic lag inflation or compression. This curriculum teaches the encoder to operate in environments ranging from nearly synchronous to strongly delayed. It also improves robustness when the model is tested on a site whose release behavior is more extreme than what appeared in training [24].

The architecture can be summarized as learning three distinct objects from the same sequence. It learns local multimodal syntax through masking, shared encounter semantics through paired contrast, and forecast-relevant anticipation through supervised horizons. The benchmark allows these components to be tested separately, because the same encounter can be replayed with or without release delay. That replayability is what makes clock-consistent representation learning empirically measurable rather than only conceptually appealing.

4. Training Procedure and Comparative Systems

Model comparison in this setting can be misleading if the baselines are weak, since almost any sequence model looks good against lagged tabular predictors built from a small set of chart fields. The experimental design therefore includes a deliberately strong comparator suite. Five systems are trained under identical encounter splits and horizon definitions. The first is a gradient-boosted tree ensemble on release-view tabular snapshots, which approximates the standard chart-time forecasting pipeline used in many retrospective studies. The second is a tabular transformer operating on acquisition-time structured fields only, thereby isolating the benefit of earlier visibility within the structured subset. The third is a sequence mixer consuming all modalities but without paired clock regularization. The fourth is a multimodal transformer with standard contrastive pretraining but no release-aware projection. The fifth is the proposed dual-clock architecture.

All systems are evaluated under leave-one-site-out rotation. In each fold, seven sites are used for training, one for validation, and one for held-out evaluation. The validation site rotates independently from the test site so model selection does not overfit to the same domain used for final reporting. Within each training fold, encounter identities are split chronologically to prevent prefix leakage through repeated windows from the same encounter. This matters because rolling windows can inflate performance if one trajectory contributes highly similar prefixes to both training and evaluation. The chronology-based split is stricter than random window splitting and better reflects deployment.

Class imbalance is substantial because the escalation endpoint is uncommon at the encounter level and even more uncommon at the per-prefix level [19]. Positive prefix rate at the 6-hour horizon is 8.7% on average, with fold-specific variation from 6.1% to 11.3%. Rather than global oversampling alone, the training procedure uses a horizon-weighted sampler that increases the frequency of prefixes farther from onset. Without this weighting, the model tends to focus on easy near-onset positives and yields high area under the receiver operating curve but poor early ranking. The weighting scheme is smooth rather than binary, gradually emphasizing prefixes whose onset is still distant but not too far to be semantically unrelated.

Optimization uses AdamW with linear warmup followed by cosine decay. The backbone dimension is fixed at 320 for all

transformer-based models to ensure that performance differences are not explained by parameter count alone. The proposed model contains 28.6 million trainable parameters, compared with 27.9 million for the standard multimodal transformer and 24.1 million for the sequence mixer. Masking ratio is 0.22 during pretraining and remains active at a lower rate during supervised finetuning. The paired-view temperature parameter is tuned on the validation site. Early stopping relies on a composite validation score averaging precision-recall area, calibration error, and forecast advance interval on the validation fold.

The benchmark allows a natural definition of advance interval. For a threshold θ, let $\tau_i^A(\theta)$ be the first time the acquisition-view model crosses threshold for encounter i, and let $\tau_i^R(\theta)$ be the analogous time for the release-view baseline. The main timing metric is

$$\Delta_i(\theta) = \tau_i^R(\theta) - \tau_i^A(\theta),$$
$$\bar{\Delta}(\theta) = \text{median}_{i:\tau_i^A<\infty} \Delta_i(\theta). \tag{11}$$

A positive value indicates that the acquisition-aware model raised its forecast earlier. This definition differs from measuring simple time-to-onset. It compares models directly on the same encounter and same threshold scale, thereby isolating relative timing rather than absolute earliness.

Calibration is evaluated through expected calibration error and threshold-conditioned reliability diagrams, but the training process also uses differentiable binning to penalize miscalibration. Forecasts in this domain can be ranked well while still being operationally poor if the probability scale is distorted. Since the benchmark spans heterogeneous sites, a model that is well calibrated only on average but poorly calibrated in specific domains is not acceptable. The calibration term is therefore computed per site and then averaged, with an extra penalty on between-site variance [22].

To test whether gains arise from multimodality or from clock handling, three intervention experiments are built into training and evaluation. The first intervention is release forcing, where selected modalities are delayed from acquisition to release time. The second is content coarsening, where continuous monitor summaries are replaced by sparsified chart-style aggregates and note embeddings are averaged into batch-level vectors. The third is order jitter, where event timestamps are randomly perturbed inside local windows. These interventions are applied both to the proposed model and to the multimodal transformer baseline. If the proposed method merely benefits from larger capacity, it should not display qualitatively different sensitivity curves under these manipulations.

Confidence intervals are computed by encounter-level bootstrap with 1,000 resamples per fold. Since rolling windows from the same encounter are correlated, window-level bootstrap would produce spuriously narrow intervals. Statistical comparison between models uses paired bootstrap on per-encounter average losses and threshold metrics. The study is not framed as a null-hypothesis exercise, but the intervals help distinguish systematic timing gains from chance variation under highly imbalanced labels.

One practical complication is that the release-view baselines sometimes see an abrupt accumulation of charted values when delayed entries are batch-published. This can create an artificial surge in apparent predictive power very near onset. To avoid exaggerating the strength of such late bursts, the analysis separately reports ranking performance in the final hour before onset and in the earlier portion of the 6-hour window. Models that derive most of their success from last-minute batch releases are easy to identify under this breakdown.

The overall training protocol is therefore designed to answer several layered questions at once. Which model ranks future events best under external-site evaluation. Which one remains calibrated when documentation behavior changes. Which one raises forecasts earlier relative to chart-constrained baselines. Which one is most sensitive to missing modalities or altered lag kernels. The benchmark’ s replay structure makes these questions comparable across identical encounters, which is rare in multimodal healthcare forecasting and essential for the claims made later in the paper [10].

5. Empirical Results

The first result is that the strongest gains appear in precision-recall space rather than in receiver-operating performance alone. The release-view gradient-boosted baseline reaches a mean event-level area under the precision-recall curve of 0.284 across the nine held-out sites. The acquisition-time tabular transformer improves this to 0.332, showing that earlier visibility within structured fields is already beneficial. The sequence mixer using all modalities but no paired clock objective attains 0.381. The multimodal transformer with standard contrastive pretraining reaches 0.401. The proposed dual-clock architecture yields 0.462. The bootstrap interval for the difference between the proposed method and the standard multimodal transformer excludes zero in every fold, with a pooled mean improvement of 0.061. Receiver-operating values rise more modestly, from 0.861 for the release-view baseline to 0.914 for the proposed model, reinforcing the familiar point that rare-event sequence forecasting is better summarized by precision-recall than by receiver-operating curves.

The timing metric reveals a larger separation than discrimination alone suggests. At matched operating points yielding 0.40 precision, the median advance interval of the acquisition-aware tabular transformer over the release-view gradient-boosted baseline is 31 minutes. The sequence mixer extends this to 54 minutes. The standard multimodal transformer reaches 67 minutes. The proposed system reaches 109 minutes. The advantage is not confined to a few exceptionally early encounters. At the same precision level, 43.8% of positive encounters receive at least a 90-minute earlier forecast from the proposed model than from the release-view baseline. For the standard multimodal transformer the corresponding figure is 27.1%. These values show that the paired clock objective affects not only whether forecasts are correct but also when the model becomes confident enough to cross threshold.

A more granular look at forecast trajectories shows that release-view models often remain flat until a burst of chart availability arrives, after which the risk score rises sharply. Acquisition-aware models behave differently [6]. Their curves tend to climb in smaller increments as new signals appear, particularly from note embeddings, image-side descriptors, and device logs. This smoother accumulation produces better ranking in the middle region of the 6-hour horizon, where actionability is still plausible and where late-burst baselines

Table 1. Corpus Composition and Scale

Component	Count	Notes	Units
Encounters	58,432	Multi-site	trajectories
Forecast states	4.81M	Rolling windows	states
Observations	214.7M	Multimodal events	tokens
Sites	9	Distinct delay regimes	institutions
Modalities	5	Structured, text, image, etc.	types
Median length	19.6	Heavy-tailed distribution	hours

Table 2. Modality Families and Characteristics

Modality	Temporal Density	Delay Pattern	Representation
Structured fields	Sparse	Short lag	Tabular codes
Monitor summaries	Dense	Near real-time	Window vectors
Text embeddings	Irregular	Long-tailed lag	Sentence vectors
Image descriptors	Sparse	Workflow-dependent	Scalars
Device/workflow logs	Burst-like	Minimal/batched	Event counts

have not yet caught up. In the final hour before onset the gap narrows, which is expected because even delayed chart entries begin to reveal strong evidence. The central value of the proposed system lies farther upstream.

Calibration under site shift was one of the most demanding tests. The release-view gradient-boosted baseline produces an average expected calibration error of 0.074, with marked site-to-site variability. The acquisition-time tabular transformer lowers this to 0.058. The sequence mixer reaches 0.046. The standard multimodal transformer attains 0.041. The proposed dual-clock model reaches 0.028, while also showing the lowest between-site spread. Calibration slope averages 0.95, compared with 0.82 for the release-view baseline. Notably, the proposed model remains well behaved even on sites with lag kernels more severe than any single training site, suggesting that curriculum exposure to delay helps prevent probability inflation when observability degrades.

The role of multimodality is nuanced. Removing note embeddings from the proposed model reduces precision-recall area by 0.021 and shrinks the median advance interval by 18 minutes. Removing image-side descriptors reduces precision-recall area by 0.015 and shortens the advance interval by 12 minutes. Removing device or workflow logs produces a smaller precision-recall drop, 0.011, but a larger degradation in calibration, particularly on sites whose documentation lag is high. This indicates that device and workflow events stabilize interpretation even when they are not the main ranking signal. Removing continuous monitor summaries has the strongest impact on raw discrimination, reducing precision-recall area by 0.034 and collapsing some of the early lift. The architecture therefore does not derive its gains from one spectacular modality [13]. The improvement is distributed, with certain modalities contributing more to timing and others more to confidence control.

Release forcing provides the clearest evidence for the study′ s main claim. When all modalities in the proposed model are delayed from acquisition time to release time while preserving their content, precision-recall area falls from 0.462 to 0.356, and the median advance interval over the release-view baseline shrinks from 109 minutes to 37 minutes. If the gain had come mostly from richer content or bigger capacity, such a large drop would not be expected. The decline is not identical across modalities. Forcing only note embeddings to release time reduces the advance interval by 16 minutes. Forcing only image-side descriptors reduces it by 11 minutes. Forcing only device logs costs 9 minutes. Forcing monitor summaries has the most severe impact, costing 28 minutes. These patterns confirm that the benchmark is indeed measuring observability, not just variable count.

Content coarsening yields a different profile. When monitor summaries are replaced by chart-like aggregates while retaining acquisition time, precision-recall area falls by 0.027 and the advance interval by 22 minutes. When note embeddings are averaged into coarse batch vectors rather than event-local fragments, precision-recall area falls by 0.013. The effect is therefore real but smaller than full release forcing. This distinction matters. It means that the benchmark′ s gains are driven both by earlier access and by finer temporal granularity, yet earlier access is the dominant factor. The study thus separates two ideas that are often conflated: high-frequency representation and early visibility.

The order-jitter intervention shows that local sequencing carries predictive content beyond aggregate counts. Perturbing timestamps within 15-minute windows causes the proposed model′ s precision-recall area to decline by 0.018 and its advance interval by 14 minutes. The standard multimodal transformer degrades more sharply under the same perturbation, with a 0.029 drop in precision-recall area and a 23-minute loss in advance interval [8]. This suggests that the order-sensitive regularizer teaches the encoder to preserve meaningful sequence structure while avoiding excessive fragility to harmless reorderings. In practical terms, the model can tolerate moderate timestamp noise without forgetting the difference between informative event sequences and arbitrary event bags.

Site-specific behavior adds another layer. The best absolute performance occurs on Site A, where modality coverage is broad and lag variance is moderate. The most striking relative gains, however, appear on Sites D and G, which have long-tail release behavior for structured documentation and text. On

Table 3. Baseline and Proposed Systems

Model	Input View	Modalities	Key Property
GBDT baseline	Release	Structured	Tabular snapshot
Tabular transformer	Acquisition	Structured	Early visibility
Sequence mixer	Acquisition	All	No clock pairing
MM transformer	Acquisition	All	Contrastive only
Dual-clock model	Both	All	Clock consistency

Table 4. Main Performance Comparison

Model	PR-AUC	AUROC	Calibration Error
GBDT (release)	0.284	0.861	0.074
Tabular (acq.)	0.332	0.882	0.058
Seq. mixer	0.381	0.897	0.046
MM transformer	0.401	0.905	0.041
Dual-clock	0.462	0.914	0.028

Site D the proposed model improves precision-recall area over the release-view baseline by 0.198 and advances forecasts by a median of 142 minutes. On Site G the gain is 0.171 and 136 minutes. Site H, with sparse imaging, shows a smaller but still substantial gain of 0.143 and 82 minutes. These results indicate that the proposed method benefits most where documentation delay obscures a large portion of the early signal frontier. The model does not simply exploit sites with rich data. It helps most when the benchmark' s dual-clock structure matters most.

The prevalence perturbation experiments further probe robustness. Event prevalence is synthetically adjusted within each held-out site to 5%, 10%, and 15% while recalibrating thresholds on the validation site only. Precision-recall values move as expected, but the ordering of models does not change. More interestingly, the proposed model retains lower calibration error than all comparators at each prevalence level. The advance interval is also comparatively stable. This suggests that the dual-clock representation does not depend on one narrow event rate regime, an important property for deployment across institutions whose baseline severity differs.

One of the most revealing diagnostics concerns false positives. The release-view baseline produces a large cluster of near-onset false positives triggered by late chart bursts that look highly event-like but arise in ultimately negative encounters [27]. The proposed model shifts the false-positive mass earlier and makes it sparser. Many of its errors occur in trajectories with transient multimodal disturbance that later resolves, which is a more acceptable failure mode than being overconfident only after the chart suddenly becomes dense. In other words, the architecture trades some late brittle certainty for earlier, better-calibrated suspicion. That trade is visible in trajectory plots and aligns with the benchmark' s stated goal.

Another diagnostic involves positive encounters missed by all models. These trajectories are enriched for abrupt escalations with weak precursor structure and for sites where one or more high-value modalities are absent. The proposed model still misses some of these cases because no sequence encoder can forecast reliably when the necessary information is either unobserved or released almost simultaneously with onset. This limit is useful to state plainly. The benchmark does not imply that every event possesses a recoverable earlier-than-EHR signature. It shows that many do, and that a carefully designed model can capture them more consistently than record-constrained baselines.

6. Ablation Trajectories and Transport Stress Testing

Ablation analysis is often reported as a small appendix to main results, but in this study it is central because the chief claim concerns timing semantics rather than architecture alone. The first ablation removes the paired clock projection while keeping the rest of the proposed model unchanged. This causes precision-recall area to fall from 0.462 to 0.417 and increases expected calibration error from 0.028 to 0.037. More revealingly, the advance interval drops by 29 minutes. The model still benefits from multimodality, yet it becomes less capable of distinguishing clock-invariant encounter structure from truly earlier information. Risk curves in this ablation are steeper and noisier, indicating that the projection term serves as a stabilizer rather than only as an accuracy enhancer.

The second ablation removes masked modeling. The resulting system is supervised end-to-end on dense forecast windows without auxiliary reconstruction. Precision-recall area falls only modestly to 0.446, but robustness under modality dropout degrades sharply [20]. When note embeddings are removed at test time, the no-mask variant loses 0.034 precision-recall points, compared with 0.021 for the full model. When monitor summaries are partially corrupted, calibration deteriorates more sharply than in the full model. This suggests that masked learning primarily improves local redundancy across modalities, allowing the encoder to reconstruct latent event context even when one stream becomes sparse or noisy.

The third ablation drops the adversarial site invariance term. Internal validation performance improves slightly, but held-out site performance worsens on five of nine folds. The average precision-recall area becomes 0.451, and calibration error increases to 0.034. Hidden-state probing reveals that site identity becomes easier to decode, indicating that the encoder has learned low-level documentation patterns that do

Table 5. Advance Interval at Matched Precision

Model	Precision Level	Median Advance	Units
Tabular (acq.)	0.40	31	minutes
Seq. mixer	0.40	54	minutes
MM transformer	0.40	67	minutes
Dual-clock	0.40	109	minutes

Table 6. Modality Ablation Effects

Removed Modality	PR-AUC Drop	Advance Loss	Calibration Impact
Text embeddings	-0.021	-18 min	Mild
Image descriptors	-0.015	-12 min	Mild
Device logs	-0.011	-9 min	Moderate
Monitor summaries	-0.034	-28 min	Strong

not transport. This is an important caution because dual-clock inputs create a temptation for the model to learn one hospital′ s release habits as if they were universal markers of risk. The adversarial term does not eliminate all site information, nor should it, but it materially reduces shortcut learning.

A more severe stress test replaces the original delay kernels of the test site with kernels sampled from a different site while keeping encounter content fixed. Under this mismatch, the release-view gradient-boosted baseline is unstable, with fold-specific precision-recall changes ranging from -0.041 to $+0.018$. The proposed model changes much less, with a range from -0.012 to $+0.007$. This does not mean the proposed system is insensitive to documentation policy. Rather, it means the hidden state is less tightly coupled to one particular release convention. That property is exactly what the benchmark is meant to reward.

The next stress test addresses partial observability. At test time, one randomly chosen modality family is removed from each positive encounter and from a matched set of negatives. This is harsher than training-time dropout because the model does not know which modality will disappear. The proposed system loses 0.026 precision-recall points on average, compared with 0.041 for the standard multimodal transformer. The largest gap appears when note embeddings are absent [14]. The paired clock objective seems to encourage the model to preserve enough semantic overlap across modalities that missing text can be partially compensated by device and structured streams. This does not amount to perfect substitution, but it is empirically meaningful.

Sequence length is another source of brittleness in many event transformers. The benchmark includes trajectories whose prefix lengths vary by more than an order of magnitude. To test scalability, evaluation is stratified by prefix event count. The proposed system maintains relatively stable precision-recall performance from short prefixes through very long ones, while the standard multimodal transformer degrades more sharply at the longest lengths. The hybrid transformer-state-space backbone appears to matter here. Earlier-than-EHR forecasting often requires reading long sequences because unreleased evidence may be sparse and scattered. A method that performs well only on short, dense prefixes would be of limited value.

A particularly informative perturbation concerns release burstiness. Some sites publish many values in clusters after periods of inactivity. Synthetic burst inflation makes those clusters even more extreme without changing total content. The release-view baseline becomes overconfident under this manipulation, since batch publication creates abrupt feature pileups. The proposed model is less affected because its acquisition-view training teaches it that semantically related evidence may have been spread out earlier. In other words, the model has learned a partial deconvolution of documentation bursts. That capability is one of the practical dividends of dual-clock training.

The study also tests a counterfactual in which the release-view baseline is given perfect knowledge of future delay statistics for the test site, allowing post hoc lag-aware feature engineering. This stronger baseline recovers some calibration and modestly improves advance interval, but it remains well below the proposed system. The reason is instructive [2]. Knowing the average delay of a modality does not restore the exact order and freshness of individual observations. The benchmark′ s advantage comes from event-level visibility, not from coarse delay priors alone. This again supports the argument that observability should be encoded directly, not approximated after the fact.

Finally, the hidden-state geometry is inspected using linear probes trained to predict modality presence, site identity, and forecast horizon label. The proposed model yields higher probe accuracy for horizon labels than for site identity relative to the ablated variants. This indicates that the representation is organized more around future event structure than around documentation style. The paired projection and adversarial term jointly appear to redirect capacity toward the forecasting task rather than toward accidental domain markers. Such probes are not definitive explanations, but they provide one more piece of evidence that the architecture is doing something more principled than memorizing which hospital delays notes the longest.

7. Interpretation from a Data-Science Perspective

Several empirical themes emerge when the results are viewed through the lens of sequence learning rather than through a

Table 7. Release Forcing Impact

Setting	PR-AUC	Advance Interval	Change
Full acquisition	0.462	109 min	Baseline
All forced release	0.356	37 min	Large drop
Text forced	—	-16 min	Moderate
Image forced	—	-11 min	Moderate
Monitor forced	—	-28 min	Strong

Table 8. Ablation of Training Objectives

Ablation	PR-AUC	Calibration Error	Advance Change
Full model	0.462	0.028	Baseline
No pairing	0.417	0.037	-29 min
No masking	0.446	0.033	-12 min
No adversarial	0.451	0.034	-15 min

clinical narrative. The first is that timestamp handling is not a preprocessing nuisance. It changes the hypothesis class effectively available to the learner. Two models may consume the same nominal variables and still solve different problems if one sees those variables when they are generated and the other sees them only after publication lag. Benchmark design must therefore specify not merely feature content but information exposure. Without that specification, reported gains from multimodal models are difficult to compare, because the underlying observability frontier is hidden.

The second theme is that earlier access and richer representation are related but separable. The release forcing experiments show that much of the gain vanishes when earlier channels are delayed to record time, even though the model architecture and raw content remain unchanged. The content coarsening experiments show that granularity also matters, but less strongly than exposure timing. This separation is conceptually useful. It means that one can ask whether a new modality helps because it contains novel information, because it is available sooner, or because it preserves fine dynamics that later chart summaries erase [3]. Those are distinct scientific questions, and the benchmark allows them to be measured separately.

A third theme concerns transfer. Many multimodal models look impressive under random splits and lose much of that advantage when moved across domains. Here the strongest generalization comes from representations trained to preserve shared encounter semantics across clocks while suppressing gratuitous site identity. This suggests that dual-clock consistency is not merely an in-domain forecasting trick. It acts as a regularizer against one pervasive form of overfitting: learning how a particular institution publishes data rather than how the underlying process unfolds. Cross-site success therefore depends not only on feature richness but on disentangling the generative process from the documentation process.

Another important observation is that calibration improves when the model is allowed to represent uncertainty arising from unreleased evidence. Release-view baselines often behave overconfidently near batch updates because the record suddenly becomes dense. Acquisition-aware models can instead accumulate weaker pieces of evidence over time. When trained with explicit uncertainty and calibration losses, they produce smoother risk estimates that better match empirical frequencies. This matters for any downstream decision policy that depends on probability scale rather than just ranking. A well-ranked but badly calibrated earlier-than-EHR model would be difficult to trust operationally.

The study also shows that multimodal sequence learning benefits from auxiliary tasks that match the structure of the deployment problem. Masked reconstruction alone helps robustness, but paired-view contrast is what sharpens clock awareness. Adversarial site suppression helps transport, but only when the encoder still has access to genuine lag information. No single loss term is dominant. The strongest behavior emerges from the combination, which is consistent with the complexity of the benchmark. Earlier-than-EHR forecasting is not a one-objective problem. It involves local syntax, delayed observability, external shift, and class imbalance simultaneously [12].

From a benchmarking standpoint, the most useful by-product may be the replay methodology itself. Because each encounter can be observed under multiple visibility regimes, the benchmark enables within-encounter causal-style questions without making strong causal claims. One can ask what would have happened to model confidence if note release had been immediate, or if monitor summaries had been charted only hourly, or if local timestamp order had been noisier. These are counterfactuals about information exposure, not about patient outcomes, and they are precisely the counterfactuals needed to understand whether a forecasting model benefits from modality content or from visibility timing.

The benchmark also exposes a limit that is easy to forget in methodological enthusiasm. Some events simply do not offer a substantial anticipatory signature, at least not in the observed modalities. For those cases, no amount of clock-aware modeling will produce large advance intervals. This negative result is useful because it discourages overgeneralization. The purpose of the present framework is not to assert that every risk signal exists far upstream. It is to provide a clean way to determine when such signals exist, how much they help, and which modeling choices recover them.

Table 9. Robustness Under Perturbations

Perturbation	PR-AUC Change	Advance Change	Observation
Order jitter	-0.018	-14 min	Moderate sensitivity
Content coarsening	-0.027	-22 min	Loss of granularity
Delay mismatch	-0.012	Small	Stable transfer
Modality dropout	-0.026	Moderate	Graceful degradation

8. Discussion

The findings presented here are best understood as a statement about data representation and benchmark construction. They do not suggest that ever larger multimodal encoders will automatically solve hospital forecasting, nor do they imply that release-time charts are unusable. The more modest conclusion is that sequence models trained on record-time views leave measurable predictive value on the table when the underlying sources become machine-available earlier. Recovering that value requires both the right data abstraction and the right training objective. A benchmark lacking dual clocks cannot make the distinction. An encoder lacking clock consistency will often blur it.

One consequence is that reproducibility in multimodal forecasting should include time semantics. Releasing code and model weights is not enough if the data pipeline silently rewrites observation times. A study that claims early prediction from notes, images, or device logs should specify when those signals become visible to the model and whether that visibility matches a plausible operational setting [5]. Otherwise replication is ill-posed because the benchmark itself has changed. The dual-clock formalism used here offers one way to make such semantics explicit and portable.

A second consequence is that model comparison should move away from purely static endpoint summaries. Event-level precision-recall area is still informative, but it misses the central distinction between a model that forecasts early and a model that becomes accurate only after the record catches up. The advance interval metric is not the only possible timing measure, yet it captures a dimension of performance that ordinary ranking scores suppress. In many applications the timing dimension may be the more valuable one, especially when actions become narrower as the event approaches. Sequence forecasting benchmarks that ignore this dimension are incomplete.

The work also suggests that external evaluation under documentation shift is a necessary complement to ordinary domain shift analysis. Cross-site prediction has typically focused on differences in patient mix, prevalence, or variable definitions. Those remain important, but release behavior is another axis of drift and can be just as damaging. A model tuned to one institution′ s publication habits may degrade even when the underlying event process is similar. The adversarial and curriculum components of the proposed system helped, yet they should be viewed as initial responses rather than final solutions. Richer domain adaptation strategies that explicitly model site-specific delay kernels are a promising direction.

Several limitations should be noted. The benchmark, although empirical, is still constructed around a replay formalism and predefined endpoint logic. Upstream text and image features are frozen rather than learned jointly. Continuous monitoring is summarized into windows rather than processed as raw waveforms. These choices were made to keep the benchmark tractable and to focus the contribution on time semantics. They also mean that the reported values are best interpreted as method-comparison results within this corpus, not as universal ceilings for multimodal forecasting [7]. A second limitation is that site identity is known during training. Completely unsupervised transport to unseen release regimes remains harder than the current experiments suggest. A third limitation is that the escalation bundle, while operationally clear, does not capture all meaningful adverse trajectories equally well.

Even with those caveats, the study offers a concrete path forward. Future benchmarks can adopt dual-clock storage, release-aware evaluation, and counterfactual visibility replay as standard components. Future models can treat publication lag as a signal to be reasoned about rather than as a nuisance to be ignored. More broadly, the field can stop conflating richer modalities with better temporal access. That distinction is foundational. It determines not only how models are trained, but what progress claims actually mean.

9. Conclusion

The empirical evidence reported here supports a simple claim: in multimodal hospital forecasting, the moment when evidence becomes visible to the learning system is itself part of the problem definition. When that moment is delayed to match the official chart, a substantial portion of predictive value disappears. A benchmark that preserves both acquisition and release clocks makes this loss measurable, while a representation learner trained on paired clocks can recover much of the missing signal.

Across nine sites, strong baselines, and several controlled perturbations, the proposed dual-clock architecture consistently improved precision-recall performance, calibration, and advance interval. The most important result was not a marginal gain in a static score. It was the repeated observation that forcing early channels to wait for release time removed a large share of the improvement. That pattern identifies observability, rather than parameter count, as the main source of advantage.

The broader implication is methodological. Earlier-than-EHR modeling should be formulated as sequence learning under timestamp corruption and distribution shift. Once written that way, multimodal forecasting becomes easier to benchmark honestly, compare reproducibly, and transport across institutions. That reframing may prove as important as any specific encoder used in the present study [9].

References

[1] Mahmoud A. Abozeid, Ahmed A Abou-Zeid, Hamdy Mohammed Hussein, and Mohamed Ahmed Omar. Totally versus planned conversion laparoscopic colectomy for colorectal cancers: the beginning of the learning curve. *SVU-International Journal of Medical Sciences*, 4(2):240–245, 8 2021.

[2] Omer Aziz, A Albeyati, M Derias, N Varsani, Hutan Ashrafian, Thanos Athanasiou, Susan K. Clark, John T. Jenkins, and Robin H. Kennedy. Anastomotic leaks can be detected within 5 days following ileorectal anastomosis: a case-controlled study in patients with familial adenomatous polyposis. *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, 19(3):251–259, 3 2017.

[3] M Baeza-Murcia, G Valero-Navarro, E Pellicer-Franco, V Soria-Aledo, M Mengual-Ballester, J A Garcia-Marin, L Betoret-Benavente, and J L Aguayo-Albasini. Bundles reduce anastomosis leak in patients undergoing elective colorectal surgery. a propensity score-matched study. *Frontiers in surgery*, 10:1119236–1119236, 2 2023.

[4] Tao Bao, Kun-Kun Li, Bi Liu, Xiao-Long Zhao, Ying-Jian Wang, and Wei Guo. Learning curve and associated prognosis of minimally invasive mckeown esophagectomy. *The Annals of thoracic surgery*, 114(3):933–939, 2 2022.

[5] W Bohle, Ioannis Louris, Andre Schaudt, Joerg Koeninger, and Wolfram G. Zoller. Predictors for treatment failure of self-expandable metal stents for anastomotic leak after gastro-esophageal resection. *Journal of gastrointestinal and liver diseases : JGLD*, 29(2):145–149, 6 2020.

[6] R. C. A. Dancer, Dhruv Parekh, Gavin D. Perkins, and David R Thickett. S97 long term survival in patients who undergo oesophagectomy is lower in patients who develop post-operative acute respiratory distress syndrome. *Thorax*, 69(Suppl 2):A52–A53, 11 2014.

[7] Ahmed Dehal, Sunal Patel, Hanna Park, Patrick Nguyen, Robert Yuhan, and Joseph H Ruan. Robotic colorectal surgery: Our initial experience. *The American surgeon*, 82(10):907–910, 10 2016.

[8] Eleonora Farinella, Adonis Safar, Haydar A. Nasser, Fikri Bouazza, Gabriele Liberale, Marianne Paesmans, Raphaël Maréchal, Jean-Luc Van Laethem, Alain Hendlisz, Paul Van-Houtte, and Issam El Nakadi. Salvage esophagectomy after failure of definitive radiochemotherapy for esophageal cancer. *Journal of surgical oncology*, 114(7):833–837, 10 2016.

[9] Tanweer Hussain, Abdul Jabbar, Naveed Ahmed, Naqeeb Shah, Mujahid Zulfiqar, and Anita . The comparison of hand sewn and stapled anastomoses. *Pakistan Armed Forces Medical Journal*, 72(3):1004–7, 6 2022.

[10] Rona Keay and Alexander Walker. 39 using simulation to improve surgical departmental induction for junior doctors. *International Journal of Healthcare Simulation*, 12 2021.

[11] Amit Kumar and Vinod Kumar. Single layer versus double layer intestinal anastomoses: a comparative study. *International Surgery Journal*, 7(9):2991–, 8 2020.

[12] Yang Liu, Bowen Li, and Yunwei Wei. New understanding of gut microbiota and colorectal anastomosis leak: A collaborative review of the current concepts. *Frontiers in cellular and infection microbiology*, 12:1022603–1022603, 11 2022.

[13] Ramy Mahrose and Amr A Kasem. Pulse pressure variation-based intraoperative fluid management versus traditional fluid management for colon cancer patients undergoing open mass resection and anastomosis: A randomized controlled trial. *Anesthesiology and pain medicine*, 13(4):e135659–e135659, 7 2023.

[14] Antanas Mickevičius, Pratik Sufi, and Dugal Heath. Factors predicting the occurrence of a gastrojejunal anastomosis leak following gastric bypass. *Wideochirurgia i inne techniki maloinwazyjne = Videosurgery and other miniinvasive techniques*, 9(3):436–440, 7 2014.

[15] Jonathan B. Mitchem, Caitlin Stafford, Todd D. Francone, P. L. Roberts, D. J. Schoetz, P. W. Marcello, and Rocco Ricciardi. What is the optimal management of an intra-operative air leak in a colorectal anastomosis? *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, 20(2):O39–O45, 2 2018.

[16] Jin-Wei Niu, W Ning, W Y Wang, D P Pei, F Q Meng, Z Z Liu, and D G Cai. Clinical effect of preservation of the left colonic artery in laparoscopic anterior resection for rectal cancer. *Zhonghua yi xue za zhi*, 96(44):3582–3585, 11 2016.

[17] Giuseppe Rosiello, Giorgio Gandaglia, Simone Scuderi, Francesco Barletta, Nicola Fossati, Armando Stabile, Andrea Gallina, Elio Mazzone, Luca Maria Vitale, Antony Pellegrino, Carlo Andrea Bravi, Alberto Martini, Vito Cucchiara, Pierre I. Karakiewicz, Alexandre Mottrie, Alberto Briganti, and Francesco Montorsi. Reducing the risk of postoperative complications after robot-assisted radical prostatectomy in prostate cancer patients: Results of an audit and feedback intervention following the implementation of prospective data collection. *European urology focus*, 8(2):431–437, 4 2021.

[18] Binesh Sadanandan. Data-driven risk stratification for anastomotic leak: A proof-of-concept study using electronic health records. *Journal of Data Science, Predictive Analytics, and Big Data Applications*, 9(6):1–27, 2024.

[19] Muhammad S. Sajid, Kamran Khatri, Krishna K. Singh, and Mohamed H. Sayegh. Use of staple-line reinforcement in laparoscopic gastric bypass surgery: a meta-analysis. *Surgical endoscopy*, 25(9):2884–2891, 3 2011.

[20] Juan Carlos Sebastián-Tomás, Aleix Martínez-Pérez, Elías Martínez-López, Nicola de' Angelis, Marcos Gómez Ruiz, and Eduardo García-Granero. Robotic transanal total mesorectal excision: Is the future now? *World journal of gastrointestinal surgery*, 13(8):834–847, 8 2021.

[21] S.P. Somashekhar, Kumar C. Rohit, Yethadka Ramya, Shabber Zaveri, Vijay Ahuja, Arun Kumar Namachivayam, and K. R. Ashwin. Bowel anastomosis after or before hipec: A comparative study in patients undergoing crs+hipec for peritoneal surface malignancy. *Annals of surgical oncology*, 29(1):1–10, 8 2021.

[22] Doug Speake, L. Dvorkin, C. J. Vaizey, and Gordon L Carlson. Management of colonic complications of type iv ehlers-danlos syndrome: a systematic review and evidence-based management strategy. *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, 22(2):129–135, 7 2019.

[23] Man Hon Tang, Joel Shi Hao Wong, Clement Luck Khng Chia, and Daniel Jin Keat Lee. Meta-analysis on surgical management of colonic injuries in trauma: to divert or to anastomose? *European journal of trauma and emergency surgery : official publication of the European Trauma Society*, 47(5):1381–1388, 1 2021.

[24] L Traeger, S Bedrikovetski, T M Nguyen, Y X Kwan, M Lewis, J W Moore, and T Sammour. The impact of preoperative sarcopenia on postoperative ileus following colorectal cancer surgery. *Techniques in coloproctology*, 27(12):1265–1274, 5 2023.

[25] Andrea Vignali and Paola De Nardi. Endoluminal vacuum-assisted therapy to treat rectal anastomotic leakage: A critical analysis. *World journal of gastroenterology*, 28(14):1394–1404, 4 2022.

[26] Yu Yang, Qi Wang, Jiajun Luo, Yue Jiang, Rui Zhou, Shilun Tong, Zhihua Wang, and Qiang Tong. Superoxide dismutase mimic, mnte-2-pyp enhances rectal anastomotic strength in rats after preoperative chemoradiotherapy. *Oxidative medicine and cellular longevity*, 2020:3509859–11, 4 2020.

[27] Tomáš Řezáč, M. Stasek, Pavel Zbořil, and Petr Špička. The role of crp in the diagnosis of postoperative complications in rectal surgery. *Polski przeglad chirurgiczny*, 93(5):1–7, 6 2021.