STRUCTURE-AWARE SET TRANSFORMERS: TEMPORAL AND VARIABLE-TYPE ATTENTION BIASES FOR ASYNCHRONOUS CLINICAL TIME SERIES

Joohyung Lee[1] **Kwanhyung Lee**[1,2] **Changhun Kim**[1] **Eunho Yang**[1,2]
[1]AITRICS
[2]Korea Advanced Institute of Science and Technology (KAIST)

ABSTRACT

Electronic health records (EHR) are irregular, asynchronous multivariate time series. As time-series foundation models increasingly tokenize events rather than discretizing time, the input layout becomes a key design choice. Grids expose time$\times$variable structure but require imputation or missingness masks, risking error or sampling-policy shortcuts. Point-set tokenization avoids discretization but loses within-variable trajectories and time-local cross-variable context (Fig. 1). We restore these priors in **ST**ructure-**A**wa**R**e (**STAR**) **Set Transformer** by adding parameter-efficient soft attention biases: a temporal locality penalty $-|\Delta t|/\tau$ with learnable timescales and a variable-type affinity B_{s_i,s_j} from a learned feature-compatibility matrix. We benchmark 10 depth-wise fusion schedules (Fig. 2). On three ICU prediction tasks, STAR-Set achieves AUC/APR of 0.7158/0.0026 (CPR), 0.9164/0.2033 (mortality), and 0.8373/0.1258 (vasopressor use), outperforming regular-grid, event-time grid, and prior set baselines. Learned τ and B provide interpretable summaries of temporal context and variable interactions, offering a practical plug-in for context-informed time-series models.

Track: Research

1 INTRODUCTION

Electronic health records (EHR) are irregularly sampled, asynchronous multivariate time series. Unlike images or text, there is no canonical discretization of time; consequently the *input layout* presented to a neural encoder is a key design choice. Common layouts include (i) discretized *regular grids* (Lipton et al., 2015; Labach et al., 2023; Yu et al., 2024), (ii) *event-time* (sparse/irregular) grids (Che et al., 2018; Chowdhury et al., 2023), and (iii) *point-set* tokenizations over observed measurement events (Horn et al., 2020; Tipirneni & Reddy, 2022). Unless stated otherwise, we use *representation* to denote this input layout.

In a *regular-grid* layout, time is binned into fixed-width intervals (e.g., hourly) and measurements are aggregated per bin, yielding a dense tensor in $\mathbb{R}^{T\times V}$ (Lipton et al., 2015). Because most variables are unobserved in most bins, grid models typically require a filling step, ranging from heuristic imputation (e.g., LOCF) to learned interpolation/imputation modules (e.g., mTAN) (Lipton et al., 2015; Shukla & Marlin, 2021). The *event-time* grid instead indexes time by the union of observed timestamps τ and stores values in a sparse $\mathbb{R}^{|\tau|\times V}$ array; however, asynchrony still produces substantial missingness, motivating masks, time-gap features, or additional imputation (Che et al., 2018).

Attention-based sequence models (Vaswani et al., 2017) further popularized event-time layouts because they can incorporate irregular timestamps via positional/continuous-time encodings and ignore padding through masking. Yet missingness is a double-edged signal in EHR: observation patterns reflect both physiology and care processes, so models can over-rely on missingness indicators and learn sampling-policy shortcuts that fail under domain shift (e.g., different measurement practices across hospitals) (Che et al., 2018).

Point-set representations avoid discretization by treating each observed event as a token (e.g., value, time, and variable identity) and applying permutation-invariant encoders (Horn et al., 2020). This

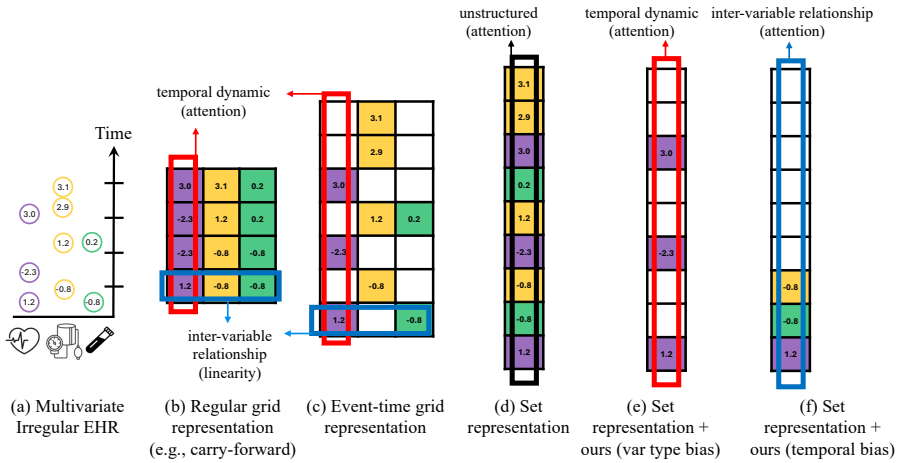

Figure 1: **EHR input layouts and biasing set attention.** (a) Irregular, asynchronous EHR events. Grid and sparse time×variable layouts (b,c) make within-variable trajectories (red) and time-local cross-variable relations (blue) explicit (sparse relies on missingness masks), whereas set tokenization (d) obscures both axes. We restore these inductive priors with a *variable-type bias* (e), favoring same-variable interactions, and a *temporal bias* (f), favoring temporally proximal interactions.

sidesteps populating unobserved grid entries and reduces dependence on heuristic imputation, but it also removes the two structural axes grids provide by construction: within-variable trajectories ("columns") and contemporaneous cross-variable relations ("rows"). As illustrated in Figure 1, these inductive biases become implicit, forcing attention to recover them from data alone.

We address this gap by augmenting point-set EHR encoders with simple, parameter-efficient *attention biases* that re-introduce grid-like structure while preserving set flexibility. We propose (i) a *variable-type bias* that encourages interactions among tokens of the same measurement type and (ii) a *temporal bias* that favors interactions among temporally nearby tokens (Figure 1). We then study where to inject these biases across Transformer depth and identify an effective layer-fusion strategy via systematic ablations over bias type and insertion schedule (Figure 2). More broadly, our approach echoes factorization choices in video modeling, where separating temporal dynamics from channel interactions improves efficiency (Tran et al., 2018).

Our contributions are:

1. **Biasing set attention for irregular EHR:** We introduce **STAR-Set Transformer** (vt–vt in Figure 2), which augments point-set EHR encoders with simple *additive attention biases* to recover grid-like inductive structure without discretization.
2. **Two complementary, parameter-efficient biases:** We combine (i) a **temporal bias** that favors temporally proximal interactions via a learnable time-distance penalty, and (ii) a **variable-type bias** parameterized by a learnable type-compatibility matrix.
3. **Systematic layer-wise ablations and consistent gains:** We evaluate bias injection schedules across transformer depth and identify effective layer-fusion strategies, yielding consistent improvements on multiple ICU prediction tasks over grid and set-based baselines.

2 Method

2.1 Set representation and its embedder for EHR

We represent an EHR episode as an irregular, asynchronous *set* of events (Horn et al., 2020). Given a mini-batch of size B, we pack episodes into a tensor $\mathbf{X} \in \mathbb{R}^{B \times L \times 3}$, where each event is a triplet

$$\mathbf{X}_{b,i,:} = \big(t_{b,i},\, v_{b,i},\, s_{b,i}\big), \tag{1}$$

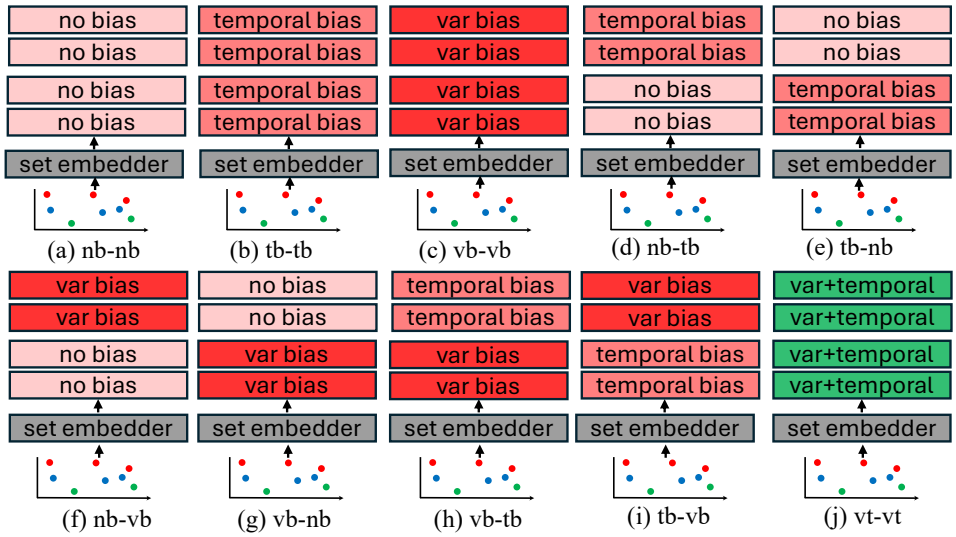

Figure 2: **Layer-wise fusion strategies for soft attention biases in the set encoder.** Each panel illustrates a bias schedule applied across Transformer encoder layers (stacked blocks from early/lower to late/upper) on top of the set embedder. We ablate *no bias* (nb), *temporal bias* (tb), *variable-type bias* (vb), and their combination (vt). The shorthand "x–y" denotes using bias x in the lower layers and bias y in the upper layers: (a) nb–nb, (b) tb–tb, (c) vb–vb, (d) nb–tb, (e) tb–nb, (f) nb–vb, (g) vb–nb, (h) vb–tb, (i) tb–vb, and (j) vt–vt. We denote vt–vt as our proposed STAR Set Transformer.

with timestamp $t_{b,i} \in \mathbb{R}$, observed value $v_{b,i} \in \mathbb{R}$, and variable/type index $s_{b,i} \in \{0, \ldots, F-1\}$. Each episode has an effective length $\ell_b \leq L$ (`input_lengths`$\in \mathbb{N}^B$); indices $i > \ell_b$ are padding. We use ITE (Tipirneni & Reddy, 2022) as the set embedder and denote its output token sequence by $\boldsymbol{H}^{(0)} \in \mathbb{R}^{B \times S \times D}$ with $S = L + 2$ (two special tokens plus L event tokens). The two special tokens are a learned `[CLS]` and a demographic token (`demo`) from a Linear–Tanh projection of age/sex.

2.2 Transformer encoder with soft attention biases

STAR-Set Transformer encodes $\boldsymbol{H}^{(0)}$ with an L_{enc}-layer (4 in this study) pre-norm Transformer encoder (H heads, GELU activations). Our core modification is to add *soft additive* biases to the attention logits, injecting (i) a temporal locality prior and (ii) a variable-type prior that are explicit in grid/sparse layouts but not in set tokenizations (Figure 1). See Appendix A.1 for more details.

Temporal bias. For each layer ℓ and head h, we learn a timescale and define an additive time-distance penalty between query token i and key token j:

$$b_{\text{time}}^{(\ell,h)}(b,i,j) \;=\; -\frac{\left|\tilde{t}_{b,i} - \tilde{t}_{b,j}\right|}{\tau_{\ell,h}}, \quad \tau_{\ell,h} = \exp(\omega_{\ell,h}) + \varepsilon. \tag{2}$$

Variable-type bias. For each layer ℓ and head h, we learn a type affinity matrix $\boldsymbol{B}_{\ell,h} \in \mathbb{R}^{F \times F}$. The variable-type bias is

$$b_{\text{var}}^{(\ell,h)}(b,i,j) \;=\; \begin{cases} (\boldsymbol{B}_{\ell,h})_{\tilde{s}_{b,i},\,\tilde{s}_{b,j}}, & \tilde{s}_{b,i}, \tilde{s}_{b,j} \in \{0, \ldots, F-1\}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Biased attention logits. Let $d_h = D/H$, $\boldsymbol{Q}^{(\ell,h)}, \boldsymbol{K}^{(\ell,h)}, \boldsymbol{V}^{(\ell,h)}$ be the standard projections, and $m_{b,j}^{\text{pad}}$ be the key padding mask (Appendix A). We modify the attention logits as

$$\begin{aligned} A_{b,i,j}^{(\ell,h)} = \frac{\langle \boldsymbol{q}_{b,i}^{(\ell,h)}, \boldsymbol{k}_{b,j}^{(\ell,h)} \rangle}{\sqrt{d_h}} \;&+\; \mathbf{1}[\texttt{bias_type}[\ell] \in \{\texttt{tb}, \texttt{vtb}\}]\; b_{\text{time}}^{(\ell,h)}(b,i,j) \\ &+\; \mathbf{1}[\texttt{bias_type}[\ell] \in \{\texttt{vb}, \texttt{vtb}\}]\; b_{\text{var}}^{(\ell,h)}(b,i,j) \;+\; m_{b,j}^{\text{pad}}. \end{aligned} \tag{4}$$

Table 1: Downstream prediction performance (AUC and APR) on three clinical tasks: CPR, mortality, and vasopressor use.

	CPR		Mortality		Vasopressor	
Model	**AUC**	**APR**	**AUC**	**APR**	**AUC**	**APR**
SMART (regular-grid) (Yu et al., 2024)	0.5098	0.0005	0.8457	0.0941	0.8239	0.0964
DueTT (regular-grid) (Labach et al., 2023)	0.6478	0.0010	0.8967	0.1527	0.8255	0.0929
PrimeNet (event-time grid) (Chowdhury et al., 2023)	0.5492	0.0006	0.6289	0.0103	0.6956	0.025
STraTS (set) (Tipirneni & Reddy, 2022)	0.5397	0.0018	0.8778	0.1457	0.8241	0.1109
STAR Set Transformer (Ours)	**0.7158**	**0.0026**	**0.9164**	**0.2033**	**0.8373**	**0.1258**

We then apply the usual $\mathrm{softmax}$ over j and compute the attention-weighted sum of values.

Layer-wise bias schedule. At each encoder layer ℓ we choose

$$\texttt{bias_type}[\ell] \in \{\texttt{nb}, \texttt{tb}, \texttt{vb}, \texttt{vtb}\}, \qquad \ell = 1, \ldots, L_{\text{enc}}, \tag{5}$$

where `nb` applies no bias, `tb` applies temporal bias only, `vb` applies variable-type bias only, and `vtb` applies both. In our ablations we often use two-stage schedules of the form "`a-b`", meaning `a` in the lower layers and `b` in the upper layers (Figure 2).

Prediction and supervised objective. After the final encoder layer and layer normalization, we use the `[CLS]` embedding as the episode representation $\boldsymbol{z}_b = \boldsymbol{H}_{b,0,:}^{(L_{\text{enc}})} \in \mathbb{R}^D$ and predict $\hat{y}_b = \boldsymbol{w}^\top \boldsymbol{z}_b + c$. For binary outcomes $y_b \in \{0, 1\}$, we train end-to-end with the sigmoid cross-entropy loss (BCE with logits):

$$\mathcal{L} = \frac{1}{B} \sum_{b=1}^{B} \mathrm{BCEWithLogits}(\hat{y}_b,\ y_b)\,. \tag{6}$$

3 Experiments

3.1 Experimental settings

Please refer to Appendix A for the details about the dataset and implementation. Note that we set our experiments to figure out where and which attention priors should be injected within a framework of a 4-layer Transformer Encoder. Details are described in Section B

3.2 Results

Table 1 compares STAR-Set Transformer to SMART and DueTT (regular-grid), PrimeNet (event-time grid), and STraTS (set) on CPR, Mortality, and Vasopressor prediction. STAR-Set Transformer achieves the best AUC and APR on all tasks: CPR reaches 0.7158 AUC (vs. 0.6478, DueTT) and 0.0026 APR (vs. 0.0018, STraTS); Mortality reaches 0.9164/0.2033 (vs. 0.8778/0.1457, STraTS); and Vasopressor reaches 0.8373 AUC (vs. 0.8255, DueTT) and 0.1258 APR (vs. 0.1109, STraTS). These gains suggest that our set-layout model effectively recovers grid-like inductive structure without committing to a fixed grid. Please refer to Appendix B for more detailed experimental results.

4 Discussion and Conclusion

Table 1 shows that **STAR Set Transformer** achieves the best results on all reported tasks and metrics. It attains AUC/APR of 0.7158/0.0026 on CPR, 0.9164/0.2033 on Mortality, and 0.8373/0.1258 on Vasopressor. Relative to the strongest baselines, this corresponds to gains of +0.0680 AUC and +0.0008 APR on CPR, +0.0386 AUC and +0.0576 APR on Mortality, and +0.0118 AUC and +0.0149 APR on Vasopressor. Overall, STAR Set Transformer delivers consistent improvements over regular-grid, event-time-grid, and prior set-based approaches on these ICU prediction tasks.

REFERENCES

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Ranak Roy Chowdhury, Jiacheng Li, Xiyuan Zhang, Dezhi Hong, Rajesh K Gupta, and Jingbo Shang. Primenet: Pre-training for irregular multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7184–7192, 2023.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pp. 4353–4363. PMLR, 2020.

Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G Krishnan. Duett: dual event time transformer for electronic health records. In *Machine Learning for Healthcare Conference*, pp. 403–422. PMLR, 2023.

Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. In *Proceedings of MLHC*, pp. 423–442. PMLR, 2023. URL `https://proceedings.mlr.press/v219/lee23a.html`.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhihao Yu, Chu Xu, Yujie Jin, Yasha Wang, and Junfeng Zhao. Smart: Towards pre-trained missing-aware model for patient health status prediction. *Advances in Neural Information Processing Systems*, 37:63986–64009, 2024.

A Dataset and Implementation Details

A.1 Dataset and Cohort Construction

We use MIMIC-IV ICU stays and follow the preprocessing protocol of Lee et al. (2023). We evaluate three prediction tasks: CPR, mortality, and vasopressor need. Unless otherwise stated, we split patients into train/validation/test sets with no patient overlap. Models are selected by the best validation AUROC and evaluated once on the held-out test set.

Task definitions. For each task, we define an observation window and a prediction horizon following Lee et al. (2023). Each prediction time requires at least 40 past observations within a 48-hour window and 20 future observations within a 12-hour window.

- **CPR:** observation window = 48 hours (min 40 events), prediction horizon = 12 hours.
- **Vasopressor:** observation window = 48 hours (min 40 events), prediction horizon = 12 hours.
- **Mortality:** observation window = 48 hours (min 40 events), prediction horizon = 12 hours.

Cohort statistics. Table 2 reports cohort sizes and class balance for each task.

Per-token metadata. Biases depend on each token's timestamp and variable ID. For event tokens we use their original $(t_{b,i}, s_{b,i})$. For the two special tokens, we assign a reference time

$$t_b^{\text{ref}} = \max_{1 \le i \le \ell_b} t_{b,i}, \tag{7}$$

with fallback $t_b^{\text{ref}} = 0$ if $\ell_b = 0$, and set their variable IDs to -1 so that type bias is disabled. Collecting metadata over the full sequence ($S = L + 2$),

$$\tilde{t}_b = \left[t_b^{\text{ref}},\ t_b^{\text{ref}},\ t_{b,1}, \ldots, t_{b,L}\right] \in \mathbb{R}^S, \qquad \tilde{s}_b = \left[-1,\ -1,\ s_{b,1}, \ldots, s_{b,L}\right] \in \mathbb{Z}^S. \tag{8}$$

Padding mask. Including the two special tokens, episode b has $\ell_b + 2$ valid tokens. We use an *additive* key padding mask $m_{b,j}^{\text{pad}}$ with $m_{b,j}^{\text{pad}} = 0$ for valid keys and $m_{b,j}^{\text{pad}} = -\infty$ for padded keys.

B Ablation Study

Goal. Figure 2 evaluates *where* and *which* attention priors should be injected when encoding point-set EHR tokens. In grid/event-time layouts, the time$\times$variable axes implicitly encourage (i) within-variable temporal modeling and (ii) time-local cross-variable context. In a set layout, these priors are not explicit; thus we ablate (a) the *bias type* (temporal vs. variable-type) and (b) the *insertion depth* (early vs. late layers) while keeping the backbone fixed.

Backbone held constant. All configurations share the same set embedder (ITE triplet embedding (Tipirneni & Reddy, 2022) with `[CLS]` and `demo` and the same L_{enc}=4-layer pre-norm Transformer encoder (Section 2). The only change across ablations is the per-layer additive bias term in the self-attention logits (Eq. 4). For any layer assigned `nb`, the bias contribution is omitted (equivalently, the additive mask is identically zero), so attention reduces to standard content-based dot-product attention with padding masking.

B.1 Bias families

We consider four mutually exclusive layer-level settings:

- `nb`: **no bias** (baseline set encoder).
- `tb`: **temporal bias** only, using the linear time-distance penalty $b_{\text{time}}^{(\ell,h)}(i,j) = -|\Delta t|/\tau_{\ell,h}$ (Eq. 2).

Table 2: Cohort statistics (MIMIC-IV).

Task	Split	# Patients / Stays	# Pos / # Neg
CPR	Train	29,544 / 37,156	561 / 1,578,395
	Val	6,341 / 8,160	54 / 31,345
	Test	6,292 / 7,934	151 / 326,508
Vasopressor	Train	29,544 / 37,156	19,023 / 1,559,933
	Val	6,341 / 8,160	1,949 / 30,739
	Test	6,292 / 7,934	3,800 / 322,859
Mortality	Train	29,544 / 37,156	7,705 / 1,571,251
	Val	6,341 / 8,160	751 / 31,304
	Test	6,292 / 7,934	1,706 / 324,953

Table 3: Layer-wise bias schedules for Fig. 2 with L_{enc}=4. Each row specifies the bias applied at encoder layers 1–4 (early→late).

Schedule	Layer 1	Layer 2	Layer 3	Layer 4
nb-nb	nb	nb	nb	nb
tb-tb	tb	tb	tb	tb
vb-vb	vb	vb	vb	vb
nb-tb	nb	nb	tb	tb
tb-nb	tb	tb	nb	nb
nb-vb	nb	nb	vb	vb
vb-nb	vb	vb	nb	nb
vb-tb	vb	vb	tb	tb
tb-vb	tb	tb	vb	vb
vt-vt	vt	vt	vt	vt

- vb: **variable-type bias** only, using the learned type affinity $b_{\text{var}}^{(\ell,h)}(i,j) = B_{\ell,h}[s_i, s_j]$ for observation–observation pairs (Eq. 3).
- vt: **joint bias** (vb+tb), adding both terms to the logits.

Bias parameters are head- and layer-specific (learned $\tau_{\ell,h}$ and $B_{\ell,h}$), but the *schedule* below determines in which layers these parameters are active.

B.2 Depth-wise fusion schedules

To isolate depth effects with minimal combinatorial growth, we use two-stage schedules of the form a-b, where a is applied to the *lower* half of layers and b to the *upper* half. With L_{enc}=4, this corresponds to layers $\{1, 2\}$ (early) and $\{3, 4\}$ (late). We evaluate the 10 schedules depicted in Fig. 2:

What these schedules test. The uniform settings (tb-tb, vb-vb, vt-vt) estimate the marginal value of each inductive prior when applied throughout depth. The swap schedules (nb-tb vs. tb-nb and nb-vb vs. vb-nb) test whether a given prior is more beneficial in early feature construction or late semantic integration. Finally, vb-tb and tb-vb test whether *ordering* the priors (type-aware mixing vs. temporal locality) across depth yields complementary benefits.

B.3 Layer-wise bias schedules

To study *where* to inject attention biases, we fix the encoder depth to L_{enc}=4 and evaluate a set of two-stage schedules (Table 3). We denote no bias as nb, temporal bias as tb, variable-type bias as vb, and their combination as vtb (vt in the table). A schedule "a-b" applies bias a to early layers (1–2) and b to late layers (3–4), enabling controlled ablations of bias type versus injection depth.

Table 4: Performance (AUC and APR) across tasks for different layer-wise bias schedules.

	CPR		Mortality		Vasopressor		Average	
Setting	AUC	APR	AUC	APR	AUC	APR	AUC	APR
nb-nb	0.600	0.002	0.918	0.202	0.838	0.123	0.785	0.109
tb-tb	0.753	0.003	0.908	0.182	0.836	0.118	0.832	0.101
vb-vb	0.665	0.001	0.918	0.191	0.840	0.120	0.808	0.104
nb-tb	0.668	0.005	0.914	0.189	0.834	0.124	0.805	0.106
tb-nb	0.693	0.004	0.916	0.189	0.838	0.118	0.815	0.104
nb-vb	0.657	0.003	0.918	0.199	0.834	0.121	0.803	0.107
vb-nb	0.683	0.003	0.918	0.193	0.838	0.118	0.813	0.105
vb-tb	0.699	0.006	0.919	0.184	0.836	0.116	0.818	0.102
tb-vb	0.684	0.001	0.917	0.199	0.839	0.119	0.813	0.106
vtb-vtb	0.716	0.003	0.916	0.203	0.837	0.126	0.823	0.111

Table 4 reports downstream performance. Temporal bias is the main driver of AUC gains: `tb-tb` achieves the best average AUC (0.832), largely due to a large improvement on CPR (0.600→0.753 AUC). Variable-type bias alone (`vb-vb`) yields smaller but consistent improvements over `nb-nb`. Combining both biases (`vtb-vtb`) provides the best average APR (0.111) while maintaining competitive AUC (0.823), and it attains the strongest APR on Mortality (0.203) and Vasopressor (0.126). Mixed schedules indicate that injecting biases only in part of the network is already beneficial, with a modest tendency for earlier-layer biasing (`tb-nb`, `vb-nb`) to outperform late-only variants (`nb-tb`, `nb-vb`).

B.4 Experimental protocol and reporting

For each schedule, we train the model under identical optimization, initialization, and data processing settings; only the per-layer `bias_type` list changes. We report AUROC and average precision (APR) for each downstream task, and summarize performance either (i) per task or (ii) by averaging across tasks after normalizing metrics (e.g., mean rank or mean Δ relative to `nb-nb`). To reduce variance, we recommend repeating each configuration with multiple random seeds and reporting mean$\pm$std; all comparisons should use the same validation-based early stopping criterion to avoid confounding depth schedules with training time.

C Limitations and Future Directions

Limitations.

- **Limited hyperparameter sweep for bias parametrization.** We perform only a narrow search over (i) the initialization of the variable-type affinity matrix $\boldsymbol{B}_{\ell,h}$ and (ii) the relative strength/scale of the temporal and variable-type bias terms (including τ initialization). Performance and stability may depend on these choices.
- **Single bias configuration per attention layer.** Our current design assigns one bias setting (`nb`/`tb`/`vb`/`vtb`) to an entire layer, so all heads share the same bias type. This may limit head specialization and prevent simultaneously modeling heterogeneous temporal and type-specific relations within a layer.
- **Supervised-only evaluation.** Experiments are restricted to supervised training on downstream ICU tasks. We do not yet evaluate STAR-Set under self-supervised pretraining, transfer/linear probing, or robustness settings (e.g., sampling-policy shift).

Future directions.

- **Systematic bias tuning and initialization study.** Conduct broader sweeps (or automated search) over $\boldsymbol{B}$ initialization (e.g., identity vs. low-rank/random), bias scaling/regularization, and τ schedules to characterize sensitivity and identify robust defaults.

- **Head-wise (channel-wise) bias routing.** Generalize STAR-Set to allow different heads within a layer to use different bias types or learned mixtures (e.g., gated `tb`/`vb` per head), encouraging explicit specialization for temporal vs. type-driven interactions.
- **Beyond supervised learning.** Integrate biases into self-supervised objectives (e.g., masked event modeling) and evaluate transfer, label efficiency, and domain-shift robustness across cohorts/hospitals.