# EMR-JEPA: FUTURE-FOCUSED AND POLICY-RESPONSIVE LEARNING FOR CLINICAL FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Clinical monitoring requires predicting a patient's future from sparse, irregular measurements collected under hospital-specific acquisition policies. We introduce EMR-JEPA, which learns patient-state representations by predicting future latent states rather than raw measurements. We first examine the masking policy: progressively shifting a matched masking budget from the past to the future increases mean out-of-domain frozen-probe AUROC from $0.6273$ to $0.6502$ across six-hour AKI and sepsis prediction on MIMIC-III, HiRID, and eICU. We then study the representation on which prediction operates. Fixed-count patching groups 16 consecutive measurements into adaptive-duration patient states, preserving finer resolution in densely observed periods. Because these states are defined by measurement count, their number and timing within the forecast horizon are unknown at prediction time and reflect the hospital's acquisition pattern; EMR-JEPA therefore predicts both alongside the states' latent content. Fixed-count patching alone matches fixed-duration patching in mean out-of-domain AUROC (0.650 for both), whereas adding count and time prediction raises it to 0.681 with one weighting shared across hospitals and tasks, and produces the clearest separation between local and long-range attention heads. Fine-tuned on a held-out hospital, EMR-JEPA attains the best average rank among self-supervised baselines for both sepsis and AKI, with the largest margin when target labels are scarce.

## 1 INTRODUCTION

Reliable forecasting of clinical deterioration requires representations that support prediction from the available patient history and generalize across hospitals and measurement acquisition policies. In critical care, anticipating sepsis and acute kidney injury (AKI) could provide valuable lead time for reassessment, patient prioritization, and resource planning (Singer et al., 2016; Khwaja, 2012). Risk estimates must be updated as new measurements arrive, while developing accurate models is complicated by severe class imbalance and limited labeled data at a new hospital. Self-supervised learning (SSL) offers a way to exploit unlabeled patient histories, but its effectiveness depends on how well the pretraining objective prepares the model for prospective prediction under these varying clinical and observational conditions.

Electronic medical record (EMR) trajectories consist of sparse observations collected at irregular intervals and asynchronously across variables. Existing methods handle this irregularity through different representations. Multi-Time Attention Networks (mTAN/mTAND) use continuous-time attention to map irregular observations to selected reference times (Shukla & Marlin, 2021). SeFT and STraTS instead encode sets of (time, variable, value) tuples, preserving an event-level representation (Horn et al., 2020; Tipirneni & Reddy, 2022). These approaches differ in temporal regularity and event-level fidelity, highlighting a central design choice for latent prediction: how to group observations into local patient states while preserving their timing and sampling-density information.

The observation process itself, however, carries both useful information and a potential source of distribution shift. What clinicians measure, and when, reflects patient condition and clinical decisions (Che et al., 2018; Agniel et al., 2018), motivating missingness-aware methods such as SMART (Yu et al., 2024). Yet acquisition patterns also vary across hospitals, so relying on them can compromise generalization (Van De Water et al., 2024). This creates a tension between robustness to acquisition differences and responsiveness to informative changes in monitoring. More frequent monitoring can



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

signal clinicians' concern about deterioration (Collins, 2013), motivating finer temporal resolution during densely observed periods. We therefore pursue a *policy-responsive* representation that forms patient states more frequently when measurements are denser and explicitly predicts the number and timing of future states. Our experiments show that retaining information about the acquisition process improves downstream forecasting, even when it increases sensitivity to simulated shifts in acquisition patterns.

A complementary question is how pretraining should reflect the temporal structure of clinical forecasting. Masked modeling learns representations by hiding parts of an input and predicting their values or representations from visible context. Clinical SSL methods use this principle to reconstruct temporal segments in PrimeNet (Chowdhury et al., 2023), predict observation presence and values in DuETT (Labach et al., 2023), and reconstruct latent representations in SMART (Yu et al., 2024). Although endpoints such as sepsis and AKI differ, their online forecasting tasks share a prospective structure: past observations are used to anticipate events within a horizon specified before deployment. Symmetric masking does not explicitly prioritize this future horizon. We therefore hypothesize that assigning higher masking probabilities to future states than to historical states better aligns pretraining with downstream forecasting. We examine this hypothesis while holding the expected total number of masked tokens fixed, isolating where predictive supervision is allocated. Our temporal reconstruction diagnostic shows that masking all future states yields the lowest future-state latent prediction loss on both source and held-out target domains, supporting this future-focused pretraining design.

Joint-Embedding Predictive Architectures (JEPAs) offer a natural framework for combining these ideas. A context encoder represents visible observations, and a predictor estimates embeddings of hidden targets supplied by a target encoder (LeCun et al., 2022; Assran et al., 2023). This latent prediction objective has supported effective image and video representation learning (Assran et al., 2023; Bardes et al., 2024). Because prediction occurs in representation space, it can reduce the pressure to reproduce unpredictable input-level detail. For EMRs, this motivates learning predictive patient states without reconstructing every raw measurement; we evaluate this potential benefit through perturbations to recorded measurement values.

Adapting JEPA to EMRs requires specifying both what constitutes a patient-state target and where that target lies in time. Irregularity and asynchrony make the construction of state units consequential, while prospective clinical forecasting provides a known temporal direction and prediction horizon for allocating supervision. We therefore ask two questions: (1) should JEPA masking emphasize the future horizon that downstream monitoring must predict, and (2) how should irregular observations be grouped into patient states, and what is gained by explicitly learning their future acquisition pattern? Together, these questions connect the encoding method to the observation process and the masking policy to the downstream prediction task.

We introduce **EMR-JEPA** to investigate these questions within a unified framework. We first compare symmetric and progressively asymmetric masking under a matched expected masking budget. We then compare fixed-duration and fixed-count patient-state encoders and isolate the contribution of future count/time supervision. The resulting model combines future-focused latent prediction with a hierarchical encoder that groups consecutive observations into states of adaptive temporal duration. Its predictive context is restricted to the observed past, while auxiliary heads learn the number and representative times of future states. Pretraining thus couples anticipation of future state content with explicit supervision of the acquisition pattern through which those states are formed. The pretrained hierarchical encoder is retained for downstream clinical forecasting.

Our contributions are threefold:

- **Future-focused latent prediction.** We investigate asymmetric masking that prioritizes the forecast horizon while keeping the expected total number of masked tokens fixed. Compared with symmetric masking, concentrating supervision in the future lowers future-state latent prediction loss and improves mean out-of-domain frozen-probe AUROC, although it can increase latent prediction loss on past states.
- **Policy-responsive patient-state encoding.** We combine fixed-count grouping of consecutive observations with explicit prediction of future state counts and representative times. Policy supervision improves mean out-of-domain AUROC from 0.6502 to 0.6814, despite greater sensitivity of the policy representation to simulated acquisition shifts. Density-stratified attention analyses further reveal differentiated local and long-range attention patterns across heads and layers.

- **Cross-database evaluation of clinical forecasting.** We evaluate six-hour sepsis and AKI forecasting on MIMIC-III, eICU, and HiRID under target-domain transfer and out-of-domain generalization. Additional experiments assess label efficiency, sensitivity to measurement noise, and the trade-off between forecasting performance and robustness to simulated shifts in acquisition patterns.

## 2 Related Work

### 2.1 Representing Sparse and Irregular EMRs

Clinical time-series encoders differ in how they aggregate measurements and represent the observation process. Benchmark pipelines commonly discretize records into regular intervals and impute missing values (Harutyunyan et al., 2019). GRU-D incorporates missingness indicators and elapsed times (Che et al., 2018), while SMART and DuETT encode binned observations with explicit missingness or observation-count information (Yu et al., 2024; Labach et al., 2023). mTAN uses continuous-time attention to evaluate representations at reference times (Shukla & Marlin, 2021). SeFT aggregates embeddings of timestamped events, whereas STraTS contextualizes observation triplets with Transformer attention (Horn et al., 2020; Tipirneni & Reddy, 2022). HeTVAE introduces UnTAN, whose parallel value and intensity pathways represent observed values and sparsity to support uncertainty-aware interpolation (Shukla & Marlin, 2022).

A related grouping choice appears in point-cloud learning: VoxelNet aggregates points within fixed spatial cells, whereas Point-MAE constructs patches from a fixed number of nearest neighbors (Zhou & Tuzel, 2018; Pang et al., 2022). EMR-JEPA applies a temporal grouping rule to heterogeneous clinical events, forming each local state from a fixed number of consecutive observations. The observation count per state is fixed, while its duration and the number of states within a forecast horizon vary with acquisition density. This differs from an hourly state grid, whose token slots and timestamps are determined by the horizon, and makes future state counts and representative times meaningful auxiliary targets.

### 2.2 Self-Supervised Learning for EMRs

Clinical SSL methods differ in both the content and temporal placement of their prediction targets. PrimeNet combines time-sensitive contrastive learning with reconstruction of masked temporal segments (Chowdhury et al., 2023). DuETT masks along time and variable dimensions and predicts both observation presence and measured values (Labach et al., 2023). The variational mTAN encoder–decoder learns through reconstruction and interpolation (Shukla & Marlin, 2021). STraTS explicitly adopts future-value forecasting as its pretext task, predicting measurements in a subsequent window from an observed history (Tipirneni & Reddy, 2022). SMART provides a close precedent for latent prediction in clinical SSL: it reconstructs EMA-encoder representations of masked observations while retaining missingness information (Yu et al., 2024).

EMR-JEPA combines latent prediction with a masking policy that prioritizes a prespecified forecast horizon. Its masking study holds the expected total number of targets fixed to examine the effect of their temporal allocation. In addition, fixed-count tokenization makes future state counts and times depend on the observation process. We supervise future token counts and representative timestamps alongside latent-state content, making a coarse future acquisition pattern an explicit pretraining target.

### 2.3 Joint-Embedding Predictive Architectures

JEPA methods learn by predicting target-encoder representations of hidden content from visible context (LeCun et al., 2022). I-JEPA instantiates this principle through masked image-region prediction with an EMA target encoder, and V-JEPA extends feature prediction to video (Assran et al., 2023; Bardes et al., 2024). Time-series adaptations include LaT-PFN, which combines JEPA with prior-data fitted networks and synthetic pretraining for in-context forecasting (Verdenius et al., 2024), and TS-JEPA, which predicts embeddings of uniformly masked temporal patches for univariate classification and forecasting (Ennadir et al., 2025).

Recent EHR preprints also apply latent prediction to patient trajectories. SMB-Structure combines next-token supervision with masked-future embedding prediction over serialized clinical records

(Adam et al., 2026). Clin-JEPA encodes hourly textualized observations and autoregressively predicts future state embeddings conditioned on intervention embeddings (Yang et al., 2026). EMR-JEPA operates directly on asynchronous structured measurements and uses fixed-count patient states with auxiliary count/time supervision. Its cross-database evaluation examines whether these representations support clinical forecasting under both target-domain adaptation and generalization to unseen databases.

## 3 The EMR-JEPA Framework

EMR-JEPA combines future-focused latent prediction with patient-state encoding that responds to the observation process (Figure 1). We first examine where to allocate prediction targets, then how to construct states and supervise their acquisition pattern within a Joint-Embedding Predictive Architecture (JEPA).

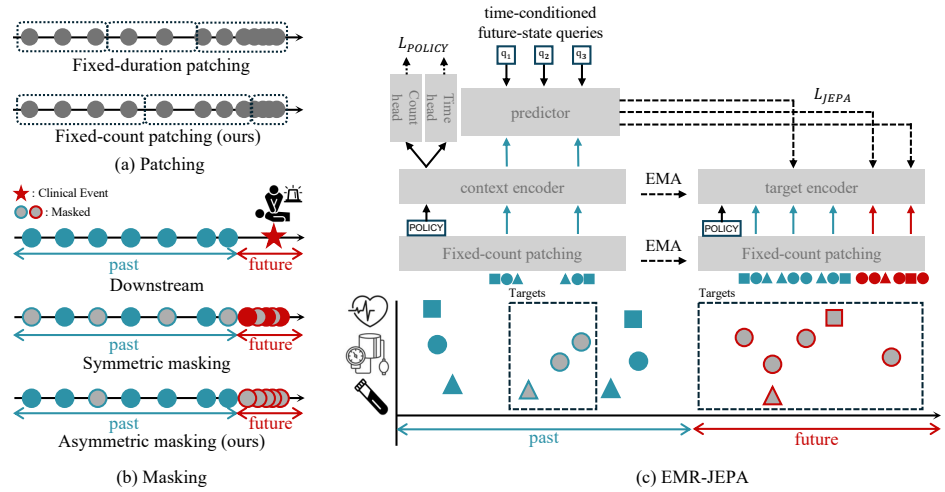



Figure 1: **EMR-JEPA overview.** (a) Fixed-count patching groups 16 consecutive measurements into patient states with adaptive temporal duration. (b) Future-focused asymmetric masking hides all future states and a subset of past states, aligning pretraining with forecasting from past observations. (c) EMR-JEPA adopts a Joint-Embedding Predictive Architecture (JEPA): a predictor uses the context encoder's representations of visible history to predict masked-state representations produced by an exponential-moving-average (EMA) target encoder. Auxiliary count and time heads predict the number and representative times of future states, providing supervision for the future acquisition pattern. The pretrained hierarchical encoder is retained for downstream clinical prediction.

### 3.1 Forecasting Setup and Design-Study Protocol

A patient stay comprises irregular observations $\mathcal{O} = \{(t_i, v_i, f_i)\}_{i=1}^{M}$, recording time, normalized value, and feature identity, together with static covariates $\mathbf{x}^{\text{sta}}$. At prediction time $\tau$, we use the preceding $H = 48$ hours to anticipate sepsis or AKI within $\Delta = 6$ hours. A *patient state* represents a local observation group and is the unit predicted during pretraining.

For the design studies, we pretrain on two source databases and fit linear classifiers to their frozen EMA-encoder representations. The third database is excluded from pretraining, classifier fitting, and selection; its test performance is OOD, whereas ID averages the source test performances. We select one SSL setting per method and fold using mean AKI/sepsis AUROC, each computed on pooled source validation data, and select probe hyperparameters separately by task. Tables 1 and 2 use one pretraining seed and one probe seed.

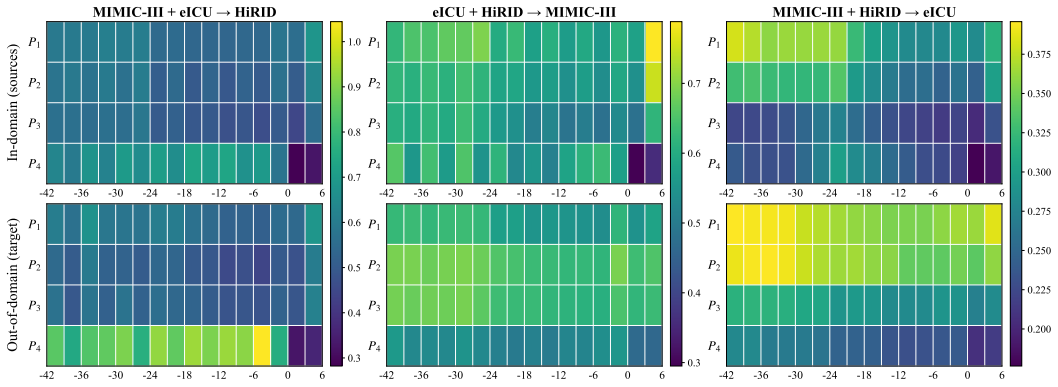



Figure 2: **Masking policies: from symmetric to asymmetric masking.** Heatmaps show reconstruction loss on source domains (top, in-domain) and held-out target domains (bottom, out-of-domain); lower is better. $P_1$ uses symmetric masking, selecting 30% of both past and future tokens, whereas $P_2$–$P_4$ progressively concentrate masking in the future, with $P_4$ masking all future tokens. All four policies mask the same expected total number of tokens within each fold: increasing future masking is offset by decreasing past masking. Time is relative to the prediction point, and 0–6 hours denotes the future horizon (right-most 2 boxes). $P_4$ achieves the lowest future-region reconstruction loss across all three splits in both evaluation regimes.

Table 1: **From symmetric to asymmetric masking policies.** Frozen linear-probe performance, with the SSL setting selected separately for each held-out fold by source-domain validation. P and F denote the past- and future-mask ratios; P ranges span the fold-specific past-mask ratios for MIMIC-III, HiRID, and eICU. Best results are in bold, with ties bolded in both rows. All results use one seed. †Shifted-axis variant: the context and target windows have equal duration, with the target window shifted forward by the forecast horizon. This is the default setting throughout the paper unless otherwise stated.

| Masking policy | | Out-of-domain | | | | | | In-domain | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AKI | | Sepsis | | Mean | | AKI | | Sepsis | | Mean | |
| | | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| $P_1$: 30.0% | F: 30.0% | 0.6362 | 0.0816 | 0.6183 | 0.0098 | 0.6273 | 0.0457 | 0.6917 | 0.1050 | 0.7579 | 0.0210 | 0.7248 | 0.0630 |
| $P_2$: 23.4–25.8% | F: 53.3% | 0.6216 | 0.0788 | 0.6572 | **0.0119** | 0.6394 | 0.0454 | 0.6906 | 0.1048 | 0.7600 | 0.0209 | 0.7253 | 0.0629 |
| $P_3$: 16.8–21.5% | F: 76.7% | 0.6307 | 0.0815 | 0.6523 | 0.0104 | 0.6415 | 0.0459 | 0.6865 | 0.1034 | 0.7546 | 0.0216 | 0.7205 | 0.0625 |
| $P_4$: 10.2–17.3% | F: 100% | **0.6511** | **0.0840** | 0.6494 | 0.0096 | **0.6502** | 0.0468 | 0.6917 | 0.1036 | 0.7616 | 0.0207 | 0.7267 | 0.0622 |
| $P_4$: 10.2–17.3% | F: 100%† | 0.6421 | 0.0821 | **0.6584** | 0.0116 | **0.6502** | **0.0469** | **0.6968** | **0.1070** | **0.7695** | **0.0221** | **0.7332** | **0.0645** |

### 3.2 From Symmetric to Future-Focused Asymmetric Masking

Initially, both views span $[\tau - H, \tau + \Delta)$, with selected patches removed from the context encoder. Policies $P_1$–$P_4$ increase future masking $r_+$ from 0.30 through approximately 0.533 and 0.767 to 1.00. Past masking $r_-$ is reduced within each fold to match the expected target budget of symmetric masking:

$$r_- \overline{L}_- + r_+ \overline{L}_+ = 0.30(\overline{L}_- + \overline{L}_+), \tag{1}$$

where $\overline{L}_-$ and $\overline{L}_+$ are mean patch counts used for calibration. Thus $P_1$ masks 30% of both regions; $P_4$ masks every future patch and fewer historical patches. Expected supervision is controlled, although realized counts vary by sample. Policies are pretrained separately without count/time supervision.

Under $P_1$–$P_3$, unmasked future measurements remain visible; $P_4$ establishes the past-to-future boundary required for forecasting. It yields the lowest future-state latent prediction loss in both future intervals across all three source–target splits, although historical loss can increase (Figure 2). Separately, mean OOD forecasting AUROC rises from 0.6273 to 0.6502, a 2.29-point gain, with non-monotonic results in some task–database settings (Table 1).

Table 2: **Patient-state construction.** Frozen-probe performance with the SSL setting selected separately for each held-out fold using source-domain validation. Fixed-Duration denotes set-per-hour tokenization; Policy denotes Fixed-Count with auxiliary count and time prediction losses. Best results are bold and underlined; second-best results are bold. All results use one seed.

| Encoding | Out-of-domain | | | | | | In-domain | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AKI | | Sepsis | | Mean | | AKI | | Sepsis | | Mean | |
| | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| Fixed-Duration | 0.6343 | 0.0816 | **0.6666** | **0.0135** | **0.6505** | **0.0476** | 0.6833 | 0.1017 | 0.7408 | **0.0223** | 0.7120 | 0.0620 |
| Fixed-Count | **0.6421** | **0.0821** | 0.6584 | 0.0116 | 0.6502 | 0.0469 | **0.6968** | **0.1070** | **0.7695** | 0.0221 | **0.7332** | **0.0645** |
| Fixed-Count + Policy | **0.6586** | **0.0938** | **0.7041** | **0.0132** | **0.6814** | **0.0535** | **0.7834** | **0.1837** | **0.8219** | **0.0311** | **0.8027** | **0.1074** |

Since $P_4$ exposes the context encoder to at most $48$ hours but the target encoder to $54$, we also test shifted, equal-duration views: context $[\tau - 48, \tau)$ and target $[\tau - 42, \tau + 6)$. Both variants attain $0.6502$ mean OOD AUROC, supporting the masking result independently of this adjustment. We use shifted views subsequently; their origins remain fixed for truncated stays.

### 3.3 Policy-Responsive Patient-State Encoding

**Fixed-count states and hierarchical encoding.** Each patch groups $p = 16$ consecutive observations, separately within past and future. Incomplete groups and observations exceeding stream caps are discarded farthest from $\tau$. Historical patch membership is shared between views; only complete patches inside the target window are eligible historical targets. The representative timestamp is the patch's last observation. Patch durations therefore adapt to acquisition density, providing finer resolution during frequent monitoring without empty bins or imputation.

A local Transformer pools time, value, and feature embeddings within each patch, following STraTS scalar embeddings (Tipirneni & Reddy, 2022). A global Transformer contextualizes these states alongside projected static covariates and, for policy supervision, a learnable `POLICY` token. Both online and EMA encoders use this hierarchy with width $d = 128$ and continuous-time RoPE in global attention. Implementation equations appear in Appendix G.

**Predicting the acquisition pattern.** Unlike hourly grids with predetermined slots, fixed-count states have an unknown future number and timing. Given maximum future-state count $N$, two heads read the contextualized policy state $\mathbf{c}_\tau = \mathbf{s}_{\text{pol}}^{\text{ctx}}$:

$$\begin{aligned} \boldsymbol{\pi} &= \text{softmax}(g_{\text{cnt}}(\mathbf{c}_\tau)) \in \mathbb{R}^{N+1}, \\ \widehat{\boldsymbol{\rho}} &= \text{sigmoid}(g_{\text{time}}(\mathbf{c}_\tau)) \in (0,1)^N, \qquad \rho_k = (t^\star_{+,k} - \tau)/\Delta. \end{aligned} \tag{2}$$

Here $\pi_c$ predicts the probability of $c$ complete future states, and $\rho_k$ is state $k$'s normalized timestamp. These targets summarize the realized acquisition pattern, not clinical decisions directly.

Table 2 separates state construction from the combined policy-supervised model. Fixed-count encoding increases mean ID AUROC from $0.7120$ to $0.7332$, with similar OOD AUROC ($0.6505$ versus $0.6502$). Adding count/time supervision raises these means to $0.8027$ and $0.6814$: a $3.12$-point OOD gain over fixed-count encoding alone.

### 3.4 Joint Latent and Acquisition-Pattern Prediction

The online encoder processes visible history bidirectionally. The EMA target encoder uses a bidirectional prefix of special tokens and historical patches, followed by causal future attention: historical targets cannot access future observations, while future targets can incorporate their own and preceding patches. The predictor attends to contextualized static and visible patch states, excluding direct access to `POLICY`. Its self- and cross-attention use continuous-time RoPE, supplemented by sinusoidal embeddings on memory patches and queries.

Observed timestamps position valid future queries during pretraining; observed counts determine valid loss slots among $N$ outputs. Count/time predictions receive auxiliary supervision. Selected

historical targets use a separate query block; attention is causal within each block and blocked between past and future queries.

Let $\widehat{\mathbf{p}}$ denote a predicted state and $\bar{\mathbf{p}}$ its EMA target. The operator $\langle\cdot\rangle_{\text{sample}}$ averages errors within each sample and then across samples with valid supervision. The objective combines latent Smooth-L1 regression, count cross-entropy, and timestamp regression:

$$\mathcal{L}_{\text{JEPA}} = \langle \text{SL1}(\widehat{\mathbf{p}} - \text{sg}[\bar{\mathbf{p}}]) \rangle_{\text{sample}}, \quad \mathcal{L}_{\text{cnt}} = -\frac{1}{B}\sum_{b} \log \pi_{b,n_b}, \quad \mathcal{L}_{\text{time}} = \left\langle (\widehat{\rho} - \rho)^2 \right\rangle_{\text{sample}},$$
$$\mathcal{L} = \mathcal{L}_{\text{JEPA}} + \lambda_{\text{cnt}}\mathcal{L}_{\text{cnt}} + \lambda_{\text{time}}\mathcal{L}_{\text{time}}. \tag{3}$$

Latent errors cover selected past and future targets; time errors cover all retained future states. Here $n_b$ is the observed future count, sg stops target gradients, and Smooth-L1 has unit transition point. We use $(\lambda_{\text{cnt}}, \lambda_{\text{time}}) = (1, 3)$ with policy supervision and omit both terms in the masking study. Target weights follow a cosine-scheduled EMA starting at momentum $0.996$; precise reductions and updates appear in Appendix G.

### 3.5 Downstream Readout and Representation Diagnostics

We initialize the downstream encoder from EMA weights, encode the complete history, and fit a linear classifier to

$$\mathbf{z}_\tau = [\text{mean}_j\, \mathbf{u}_j^{\text{ctx}}; \mathbf{s}_{\text{sta}}^{\text{ctx}}; \mathbf{s}_{\text{pol}}^{\text{ctx}}], \tag{4}$$

excluding padding from the mean and omitting the last component for models without `POLICY`. Frozen probes train only the classifier; baseline fine-tuning also updates the encoder. The predictor and auxiliary heads are discarded, so inference uses no future observations, timestamps, or counts.

In the eICU-held-out AKI ablation, the full readout outperforms both its `POLICY`-only and mean-patch-plus-static components, though `POLICY` is the least stable under test-time thinning; this readout ablation does not remove policy supervision from pretraining (Appendix H). Density-stratified attention further reveals local and long-range heads across layers (Appendix G).

## 4 Experiments

We evaluate whether future-focused latent prediction and policy-responsive patient-state encoding improve clinical forecasting across databases. We additionally examine how policy supervision affects attention patterns, representation stability under observation thinning, and the predictive information available from the `POLICY` state.

### 4.1 Evaluation Setup

Following YAIB, we harmonize MIMIC-III, eICU, and HiRID into a representation with 48 shared time-varying variables (Van De Water et al., 2024; Johnson et al., 2016; Pollard et al., 2018; Faltys et al., 2021). At each hourly prediction time, models use the preceding 48 hours to forecast sepsis or AKI onset within the next six hours. In **transfer learning**, pretraining uses two source databases, followed by supervised adaptation on the held-out target database. In **out-of-domain (OOD) generalization**, both pretraining and supervised adaptation use only the sources, and the target is excluded from training and model selection. We report AUROC and AUPRC; average ranks summarize the six target–metric combinations within each task and evaluation regime. Cohort details appear in Appendix F.

We compare EMR-JEPA with SMART, DuETT, PrimeNet, and STraTS (Yu et al., 2024; Labach et al., 2023; Chowdhury et al., 2023; Tipirneni & Reddy, 2022). The baseline comparison fine-tunes the encoder and a linear classifier. The masking, encoding, and readout studies instead use frozen pretrained encoders, while label-efficiency experiments fit target-domain linear classifiers to frozen representations. Model selection uses validation data from the permitted training domains, and seed counts are specified with each table.

EMR-JEPA uses 16 observations per patient state and latent width $d = 128$. The maximum future-state count is $N = 128$ in the MIMIC-III- and eICU-held-out folds and $N = 64$ in the HiRID-held-out

Table 3: **Cross-domain comparison with self-supervised baselines.** Results are mean ± std over three seeds. EMR-JEPA uses $(\lambda_{\text{cnt}}, \lambda_{\text{time}}) = (1, 3)$ with supervised learning rates selected by validation AUROC. Best: **bold and underlined**; second-best: underlined, based on unrounded means. Average rank is computed within each task and regime over the six target-domain–metric pairs (ties get their average rank); lower is better.

| Task | Model | eICU+HiRID → MIMIC-III | | MIMIC-III+eICU → HiRID | | MIMIC-III+HiRID → eICU | | Average Rank ↓ |
|---|---|---|---|---|---|---|---|---|
| | | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | |
| **(a) Transfer Learning** | | | | | | | | |
| Sepsis | SMART | $0.880 \pm 0.005$ | $0.018 \pm 0.002$ | $0.724 \pm 0.003$ | $0.028 \pm 0.001$ | $0.738 \pm 0.000$ | $0.020 \pm 0.000$ | 4.67 |
| | DuETT | $0.894 \pm 0.001$ | $0.026 \pm 0.002$ | $0.749 \pm 0.002$ | $0.033 \pm 0.001$ | $0.747 \pm 0.001$ | **$0.023 \pm 0.000$** | 2.25 |
| | PrimeNet | $0.891 \pm 0.002$ | $0.022 \pm 0.004$ | $0.744 \pm 0.000$ | $0.033 \pm 0.000$ | $0.722 \pm 0.001$ | $0.017 \pm 0.000$ | 4.17 |
| | STraTS | $0.897 \pm 0.001$ | $0.022 \pm 0.001$ | $0.748 \pm 0.001$ | $0.035 \pm 0.000$ | $0.744 \pm 0.002$ | $0.021 \pm 0.001$ | 2.75 |
| | **EMR-JEPA (Ours)** | **$0.914 \pm 0.001$** | **$0.031 \pm 0.002$** | **$0.769 \pm 0.001$** | **$0.038 \pm 0.001$** | **$0.749 \pm 0.002$** | $0.023 \pm 0.000$ | **1.17** |
| AKI | SMART | $0.824 \pm 0.000$ | $0.235 \pm 0.003$ | $0.745 \pm 0.001$ | $0.129 \pm 0.001$ | $0.788 \pm 0.000$ | $0.230 \pm 0.000$ | 4.00 |
| | DuETT | $0.847 \pm 0.001$ | $0.280 \pm 0.003$ | $0.778 \pm 0.001$ | $0.153 \pm 0.002$ | $0.812 \pm 0.000$ | $0.265 \pm 0.000$ | 2.25 |
| | PrimeNet | $0.807 \pm 0.000$ | $0.202 \pm 0.002$ | $0.743 \pm 0.001$ | $0.128 \pm 0.001$ | $0.785 \pm 0.000$ | $0.223 \pm 0.000$ | 5.00 |
| | STraTS | $0.840 \pm 0.001$ | $0.273 \pm 0.002$ | $0.778 \pm 0.001$ | $0.154 \pm 0.001$ | $0.799 \pm 0.000$ | $0.253 \pm 0.000$ | 2.75 |
| | **EMR-JEPA (Ours)** | **$0.850 \pm 0.000$** | **$0.288 \pm 0.002$** | **$0.820 \pm 0.000$** | **$0.205 \pm 0.004$** | **$0.839 \pm 0.000$** | **$0.286 \pm 0.001$** | **1.00** |
| **(b) Out-of-Domain Generalization** | | | | | | | | |
| Sepsis | SMART | $0.725 \pm 0.011$ | $0.006 \pm 0.001$ | $0.600 \pm 0.000$ | $0.014 \pm 0.000$ | $0.646 \pm 0.003$ | $0.010 \pm 0.000$ | 4.25 |
| | DuETT | $0.746 \pm 0.021$ | $0.007 \pm 0.001$ | $0.622 \pm 0.008$ | $0.014 \pm 0.000$ | $0.650 \pm 0.003$ | $0.010 \pm 0.000$ | 2.92 |
| | PrimeNet | $0.730 \pm 0.006$ | $0.006 \pm 0.000$ | $0.635 \pm 0.002$ | $0.015 \pm 0.000$ | $0.610 \pm 0.000$ | $0.008 \pm 0.000$ | 4.00 |
| | STraTS | $0.738 \pm 0.005$ | $0.006 \pm 0.001$ | $0.702 \pm 0.012$ | $0.022 \pm 0.002$ | $0.647 \pm 0.006$ | $0.010 \pm 0.000$ | 2.83 |
| | **EMR-JEPA (Ours)** | **$0.769 \pm 0.014$** | **$0.008 \pm 0.001$** | **$0.756 \pm 0.005$** | **$0.026 \pm 0.000$** | **$0.672 \pm 0.001$** | **$0.013 \pm 0.001$** | **1.00** |
| AKI | SMART | $0.719 \pm 0.013$ | $0.153 \pm 0.007$ | $0.657 \pm 0.010$ | $0.077 \pm 0.005$ | $0.706 \pm 0.022$ | $0.149 \pm 0.011$ | 4.33 |
| | DuETT | $0.790 \pm 0.002$ | $0.188 \pm 0.004$ | $0.697 \pm 0.010$ | $0.098 \pm 0.007$ | **$0.757 \pm 0.006$** | **$0.194 \pm 0.004$** | 2.00 |
| | PrimeNet | $0.754 \pm 0.001$ | $0.161 \pm 0.000$ | $0.690 \pm 0.001$ | $0.095 \pm 0.003$ | $0.614 \pm 0.002$ | $0.088 \pm 0.001$ | 4.00 |
| | STraTS | $0.714 \pm 0.010$ | $0.128 \pm 0.002$ | **$0.711 \pm 0.015$** | **$0.113 \pm 0.005$** | $0.716 \pm 0.003$ | $0.154 \pm 0.005$ | 3.00 |
| | **EMR-JEPA (Ours)** | **$0.793 \pm 0.000$** | **$0.211 \pm 0.004$** | $0.704 \pm 0.004$ | $0.104 \pm 0.001$ | $0.735 \pm 0.004$ | $0.164 \pm 0.002$ | **1.67** |

fold. We pretrain for 10 epochs with AdamW, a warmup–cosine learning-rate schedule, and an EMA target encoder. The policy-supervised model uses $(\lambda_{\text{cnt}}, \lambda_{\text{time}}) = (1, 3)$ across folds and tasks. Additional settings are provided in Appendix E.

### 4.2 Cross-Domain Clinical Forecasting

Table 3 compares downstream forecasting performance. Under transfer learning, EMR-JEPA achieves the highest AUROC across all six task–target settings and the highest AKI AUPRC on all three targets. It also obtains the best average rank for both tasks. Under OOD generalization, EMR-JEPA again achieves the best average rank for both sepsis and AKI. The advantage is more consistent for sepsis, where it leads both metrics on every target. For AKI, it leads on MIMIC-III, while STraTS performs better on HiRID and DuETT on eICU. These results support cross-database usefulness without implying uniform superiority in every task–domain setting. Source-domain performance and random-initialization controls are reported in Appendices D and C.

### 4.3 Temporal Allocation and Patient-State Construction

The masking study in Section 3.2 tests whether pretraining benefits from emphasizing the downstream forecast horizon. Policies $P_1$–$P_4$ fix the expected number of masked tokens while progressively shifting targets from the past to the future. Figure 2 shows that fully masking the future yields the lowest future-region loss on both source and held-out target evaluations, at the cost of degraded prediction of masked historical states. Accordingly, mean OOD frozen-probe AUROC rises from $0.6273$ under symmetric masking to $0.6502$ under full-future masking (Table 1), indicating that prioritizing future targets improves prospective forecasting under the tested budget.

Table 2 distinguishes state construction from the combined policy-supervised configuration. Fixed-count encoding alone yields similar mean OOD AUROC to fixed-duration encoding. Adding future count/time supervision raises mean OOD AUROC from $0.6502$ to $0.6814$ and mean ID AUROC from $0.7332$ to $0.8027$. This supervision is meaningful because fixed-count states have data-dependent future counts and timestamps. By contrast, an hourly grid prescribes six token slots and their grid times for a six-hour horizon. The results therefore support explicitly learning the acquisition pattern associated with observation-defined states.

### 4.4 Attention Across Temporal Scales

Figure G1 in Appendix G.4 examines the eICU-held-out checkpoint pretrained on MIMIC-III and HiRID. Fixed-count encoding yields more pronounced local attention than fixed-duration encoding, particularly in deeper layers. Policy supervision further separates attention across scales, with strongly local heads in early layers and longer-range heads in later ones. These density-dependent patterns suggest that policy supervision changes how the encoder combines nearby measurements with more distant patient states.

### 4.5 Acquisition Sensitivity and Predictive Policy Information

We test acquisition sensitivity by randomly retaining raw observations at keep rates $1.0$, $0.5$, and $0.25$. For each pretrained checkpoint, we encode the validation split of each evaluation database with and without thinning, then compare the resulting representations using cosine similarity and linear centered kernel alignment (CKA). Thinning is applied only during this diagnostic; it is not used during pretraining or supervised training.

Figure H1 in Appendix H compares fixed-count pretraining without policy supervision, the full policy-supervised representation, and its `POLICY` component. At the strongest thinning level, the full policy-supervised representation has lower cosine similarity than the model trained without policy supervision across all displayed source–evaluation combinations, and the `POLICY` component changes still more. CKA shows a similar tendency in many settings, but the comparison is not uniform: with MIMIC-III and eICU as sources, the policy-supervised representation retains higher CKA on HiRID. These diagnostics indicate sensitivity to the tested acquisition perturbation, rather than a direct reduction in forecasting accuracy under thinning.

Sensitivity to acquisition changes does not preclude predictive utility. Table H1, paired with Figure H1 in Appendix H, compares two readouts of the same policy-pretrained encoder in the eICU-held-out fold. Separate linear classifiers use either the `POLICY` state alone or the full concatenation of mean patch, static, and policy states. The `POLICY`-only readout supports both AKI and sepsis forecasting, while the full readout improves AUROC and AUPRC for both tasks in ID and OOD evaluation. Together with Table 2, these results show that acquisition-sensitive information can contribute to useful predictive representations, though this readout comparison probes one pretrained encoder rather than ablating policy supervision from pretraining.

### 4.6 Label Efficiency and Measurement Noise

We evaluate label-efficient transfer by freezing source-pretrained encoders and fitting target-domain linear classifiers on $1\%$, $5\%$, $10\%$, $50\%$, and $100\%$ of the training labels for both sepsis and AKI. EMR-JEPA achieves the highest AUROC and AUPRC at every fraction across all six task–target combinations (Table A1; Figures A1 and A2); for AKI on MIMIC-III and eICU, probes trained with $1\%$ of target labels outperform every baseline trained with $100\%$ in both metrics, supporting label-efficient reuse of the frozen representation. Robustness to measurement noise is more mixed: when Appendix B perturbs measurement values while preserving acquisition times, EMR-JEPA retains the highest sepsis AUPRC on MIMIC-III and HiRID across all noise levels, but its advantage narrows or reverses in other task–domain settings. Value-noise tolerance and sensitivity to observation thinning thus reflect distinct aspects of the learned representation.

## 5 Conclusion

We introduced EMR-JEPA, a predictive foundation model for irregular EMRs that combines fixed-count patient-state tokenization with downstream-aligned latent future prediction. By grouping consecutive observations into adaptive-duration states and predicting their future representations, counts, and times, EMR-JEPA aligns pretraining with online clinical forecasting. Across MIMIC-III, eICU, and HiRID, it achieves the best transfer-learning average rank among self-supervised baselines, though OOD gains vary by task and target. Design studies isolate the contributions of future-focused masking and joint count/time supervision beyond fixed-count tokenization, while label-efficiency, initialization, and input-corruption studies delineate its strengths and limitations.

## AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, suggesting candidate related work, and preparing LaTeX tables and figures. The authors take full responsibility for the final manuscript, including all AI-assisted content, scientific claims, and reported experimental results.

## REFERENCES


Irsyad Adam, Zekai Chen, David Laprade, Shaun Porwal, David Laub, Erik Reinertsen, Arda Pekis, and Kevin Brown. The patient is not a moving document: A world model training paradigm for longitudinal EHR. *arXiv preprint arXiv:2601.22128*, 2026. doi: 10.48550/arXiv.2601.22128. URL https://arxiv.org/abs/2601.22128.

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

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. doi: 10.1609/aaai.v37i6.25876.

Anita Collins. Effect of continuing nursing education on nurses' attitude toward and accuracy of nursing diagnosis. *International journal of nursing knowledge*, 24(3):122–128, 2013.

Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint embeddings go temporal. *arXiv preprint arXiv:2509.25449*, 2025.

Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. HiRID, a high time-resolution ICU dataset. *PhysioNet*, February 2021. doi: 10.13026/nkwc-js72. URL https://doi.org/10.13026/nkwc-js72. Version 1.1.1.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4353–4363. PMLR, 2020. URL https://proceedings.mlr.press/v119/horn20a.html.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020. doi: 10.1038/s41591-020-0789-4.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.

Arif Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184, 2012.

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 Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary Lipton, Rajesh Ranganath, Iñigo Urteaga, and Serene Yeung (eds.), *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pp. 403–422. PMLR, 2023. URL https://proceedings.mlr.press/v219/labach23a.html.

Yann LeCun et al. A path towards autonomous machine intelligence. *OpenReview*, 62(1):1–62, 2022.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision – ECCV 2022*, volume 13662 of *Lecture Notes in Computer Science*, pp. 604–621, Cham, 2022. Springer. doi: 10.1007/978-3-031-20086-1_35. URL https://doi.org/10.1007/978-3-031-20086-1_35.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Satya Narayan Shukla and Benjamin M. Marlin. Heteroscedastic temporal variational autoencoder for irregularly sampled time series. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Az7opqbQE-3.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

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*, 16(6):105, 2022. doi: 10.1145/3516367.

Robin Van De Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another icu benchmark: A flexible multi-center framework for clinical ml. In *International Conference on Learning Representations*, volume 2024, pp. 44638–44693, 2024.

Stijn Verdenius, Andrea Zerio, and Roy LM Wang. Lat-pfn: A joint embedding predictive architecture for in-context time-series forecasting. *arXiv preprint arXiv:2405.10093*, 2024.

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, et al. Clin-jepa: A multi-phase co-training framework for joint-embedding predictive pretraining on ehr patient trajectories. *arXiv preprint arXiv:2605.10840*, 2026.

Zhihao Yu, Xu Chu, Yujie Jin, Yasha Wang, and Junfeng Zhao. SMART: Towards pre-trained missing-aware model for patient health status prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63986–64009. Curran Associates, Inc., 2024. doi: 10.52202/079017-2043. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/751ef1e7f557a8a88f0837b61bf5070f-Abstract-Conference.html.

Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018. doi: 10.1109/CVPR.2018.00472. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhou_VoxelNet_End-to-End_Learning_CVPR_2018_paper.html.

# A LABEL EFFICIENCY

We evaluate transfer with frozen pretrained encoders and target-domain linear classifiers at 1%, 5%, 10%, 50%, and 100% label budgets. This protocol applies to both sepsis and AKI: it uses target labels and is distinct from source-only OOD generalization. Figures A1 and A2 visualize the numerical results in Table A1.

EMR-JEPA has the highest AUROC and AUPRC at every tested fraction across the six task–target combinations. Its 1%-label AKI probes on MIMIC-III and eICU outperform every baseline at 100% labels in both metrics. These results support label-efficient transfer of the patient-state representation, without requiring end-to-end encoder adaptation. The sweep provides point estimates, so we do not infer statistical significance from the curves.

Table A1: **Label-efficient transfer with frozen encoders.** AUROC and AUPRC are reported for varying fractions of labeled target-domain data, for both sepsis and AKI. These values correspond to Figures A1 and A2. Bold marks the best value within each task, hospital, label fraction, and metric. The supplied sweep contains point estimates without seed-wise dispersion.

| Hospital | Model | 1% | | 5% | | 10% | | 50% | | 100% | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| **Sepsis: transfer learning, frozen linear probe** | | | | | | | | | | | |
| MIMIC-III | **EMR-JEPA** | **0.8070** | **0.0097** | **0.8516** | **0.0132** | **0.8657** | **0.0146** | **0.8767** | **0.0156** | **0.8780** | **0.0163** |
| | DuETT | 0.6100 | 0.0041 | 0.7451 | 0.0083 | 0.7883 | 0.0102 | 0.8216 | 0.0128 | 0.8229 | 0.0136 |
| | STraTS | 0.7281 | 0.0064 | 0.7852 | 0.0084 | 0.7935 | 0.0087 | 0.8150 | 0.0114 | 0.8199 | 0.0127 |
| | PrimeNet | 0.7497 | 0.0066 | 0.8223 | 0.0082 | 0.8300 | 0.0085 | 0.8421 | 0.0103 | 0.8489 | 0.0113 |
| | SMART | 0.6894 | 0.0040 | 0.7665 | 0.0073 | 0.7673 | 0.0076 | 0.7978 | 0.0098 | 0.8154 | 0.0104 |
| HiRID | **EMR-JEPA** | **0.7259** | **0.0254** | **0.7677** | **0.0318** | **0.7912** | **0.0392** | **0.7909** | **0.0426** | **0.7908** | **0.0397** |
| | DuETT | 0.6571 | 0.0184 | 0.7058 | 0.0246 | 0.7031 | 0.0252 | 0.7102 | 0.0281 | 0.7141 | 0.0297 |
| | STraTS | 0.6414 | 0.0156 | 0.7009 | 0.0233 | 0.7160 | 0.0283 | 0.7287 | 0.0289 | 0.7314 | 0.0285 |
| | PrimeNet | 0.6017 | 0.0152 | 0.6769 | 0.0231 | 0.7073 | 0.0260 | 0.7168 | 0.0275 | 0.7220 | 0.0280 |
| | SMART | 0.6031 | 0.0137 | 0.6371 | 0.0166 | 0.6504 | 0.0173 | 0.6601 | 0.0182 | 0.6686 | 0.0195 |
| eICU | **EMR-JEPA** | **0.7131** | **0.0146** | **0.7353** | **0.0173** | **0.7369** | **0.0178** | **0.7458** | **0.0196** | **0.7498** | **0.0203** |
| | DuETT | 0.6567 | 0.0100 | 0.6913 | 0.0122 | 0.7029 | 0.0134 | 0.7088 | 0.0150 | 0.7101 | 0.0151 |
| | STraTS | 0.6423 | 0.0099 | 0.6845 | 0.0116 | 0.6899 | 0.0127 | 0.6999 | 0.0142 | 0.7011 | 0.0143 |
| | PrimeNet | 0.6458 | 0.0088 | 0.6756 | 0.0107 | 0.6844 | 0.0122 | 0.6937 | 0.0139 | 0.6950 | 0.0140 |
| | SMART | 0.6047 | 0.0077 | 0.6467 | 0.0090 | 0.6548 | 0.0097 | 0.6664 | 0.0103 | 0.6697 | 0.0106 |
| **AKI: transfer learning, frozen linear probe** | | | | | | | | | | | |
| MIMIC-III | **EMR-JEPA** | **0.7454** | **0.1250** | **0.7657** | **0.1416** | **0.7698** | **0.1485** | **0.7770** | **0.1562** | **0.7778** | **0.1565** |
| | DuETT | 0.6840 | 0.0944 | 0.7057 | 0.1064 | 0.7146 | 0.1102 | 0.7174 | 0.1116 | 0.7183 | 0.1138 |
| | STraTS | 0.6834 | 0.1037 | 0.7054 | 0.1106 | 0.7138 | 0.1142 | 0.7190 | 0.1156 | 0.7199 | 0.1153 |
| | PrimeNet | 0.6895 | 0.0941 | 0.7042 | 0.1068 | 0.7094 | 0.1099 | 0.7220 | 0.1172 | 0.7283 | 0.1196 |
| | SMART | 0.6301 | 0.0703 | 0.6521 | 0.0823 | 0.6587 | 0.0849 | 0.6649 | 0.0889 | 0.6658 | 0.0893 |
| HiRID | **EMR-JEPA** | **0.7285** | **0.1121** | **0.7441** | **0.1207** | **0.7502** | **0.1265** | **0.7624** | **0.1346** | **0.7638** | **0.1360** |
| | DuETT | 0.7129 | 0.1103 | 0.7252 | 0.1146 | 0.7323 | 0.1180 | 0.7413 | 0.1226 | 0.7437 | 0.1246 |
| | STraTS | 0.7136 | 0.1044 | 0.7326 | 0.1170 | 0.7340 | 0.1178 | 0.7439 | 0.1276 | 0.7435 | 0.1264 |
| | PrimeNet | 0.7016 | 0.1024 | 0.7230 | 0.1156 | 0.7247 | 0.1162 | 0.7317 | 0.1226 | 0.7325 | 0.1235 |
| | SMART | 0.6490 | 0.0735 | 0.6680 | 0.0787 | 0.6729 | 0.0817 | 0.6792 | 0.0879 | 0.6819 | 0.0901 |
| eICU | **EMR-JEPA** | **0.6778** | **0.1176** | **0.6879** | **0.1229** | **0.6913** | **0.1246** | **0.6931** | **0.1265** | **0.6933** | **0.1265** |
| | DuETT | 0.6107 | 0.0951 | 0.6367 | 0.1047 | 0.6431 | 0.1077 | 0.6497 | 0.1104 | 0.6500 | 0.1106 |
| | STraTS | 0.6135 | 0.0945 | 0.6279 | 0.1000 | 0.6313 | 0.1016 | 0.6337 | 0.1028 | 0.6340 | 0.1028 |
| | PrimeNet | 0.6041 | 0.0919 | 0.6269 | 0.0990 | 0.6330 | 0.1012 | 0.6450 | 0.1045 | 0.6476 | 0.1053 |
| | SMART | 0.5822 | 0.0834 | 0.5986 | 0.0861 | 0.6047 | 0.0884 | 0.6148 | 0.0926 | 0.6171 | 0.0932 |

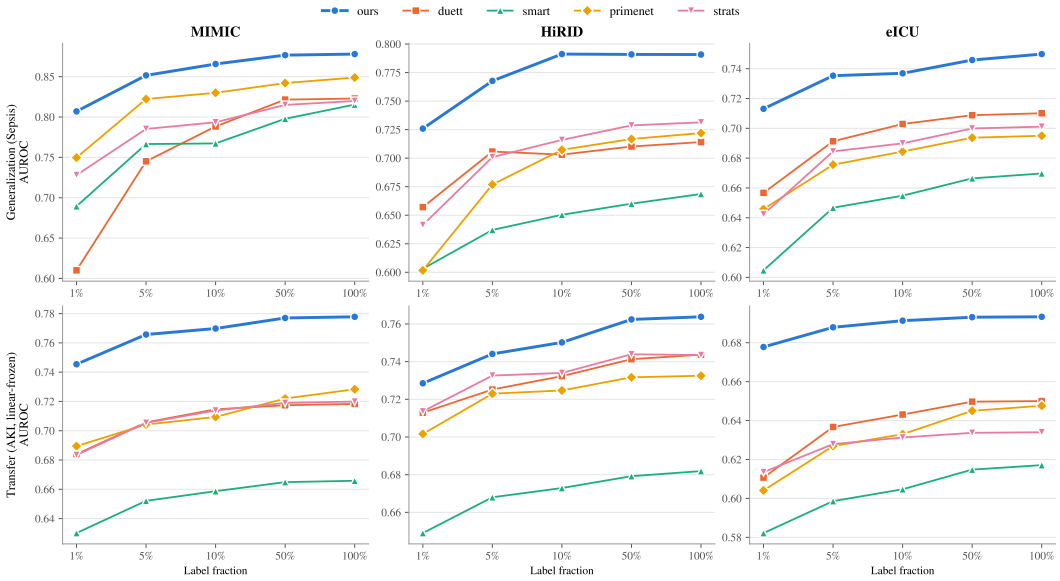



Figure A1: **Label-efficient transfer: AUROC.** Columns show MIMIC-III, HiRID, and eICU; rows show sepsis and AKI. Both tasks use frozen encoders and target-domain linear classifiers. The upper-row label "Generalization (Sepsis)" denotes this transfer experiment, not source-only OOD evaluation. Numerical values are in Table A1.

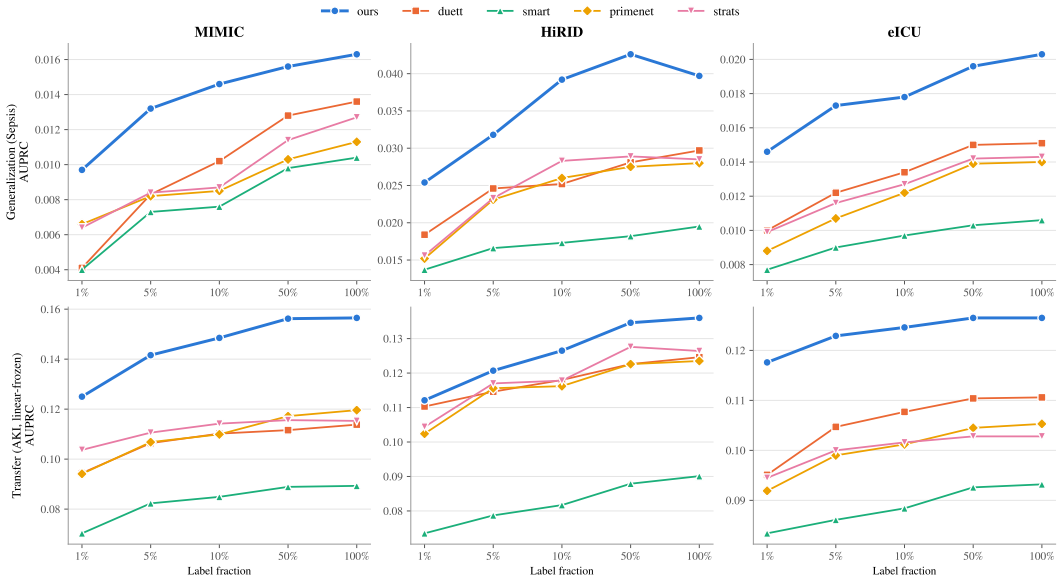



Figure A2: **Label-efficient transfer: AUPRC.** Evaluation settings match Figure A1, and numerical values are in Table A1. Both rows use target-domain frozen probes. AUPRC is interpreted within each task and hospital because prevalence differs across settings (Table F3).

## B Robustness to Additive Value Noise

**Perturbation protocol.** At evaluation, we perturb normalized observation values while leaving timestamps, variable identities, static covariates, and labels unchanged:

$$v_i' = v_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad \sigma = 10^{-S/20}, \tag{5}$$

where $S$ is the nominal SNR in decibels. The five noise levels $S \in \{26, 20, 14, 10, 5\}$ correspond to standard deviations of approximately $0.050$, $0.100$, $0.200$, $0.316$, and $0.562$ in normalized units. Lower SNR means stronger corruption. This tests sensitivity to measurement values, not changes in acquisition timing or observation density.

**Task-specific results.** Figures B1 and B2 show the six-level sweep, including the clean evaluation; Tables B1 and B2 give the corresponding values. EMR-JEPA retains the highest sepsis AUPRC on MIMIC-III and HiRID at every level, and the highest sepsis AUROC on HiRID. Its AKI advantages narrow or reverse under strong corruption, while DuETT has higher eICU sepsis AUPRC at all nonzero noise levels. Thus, useful discrimination under value noise does not imply uniformly superior robustness.

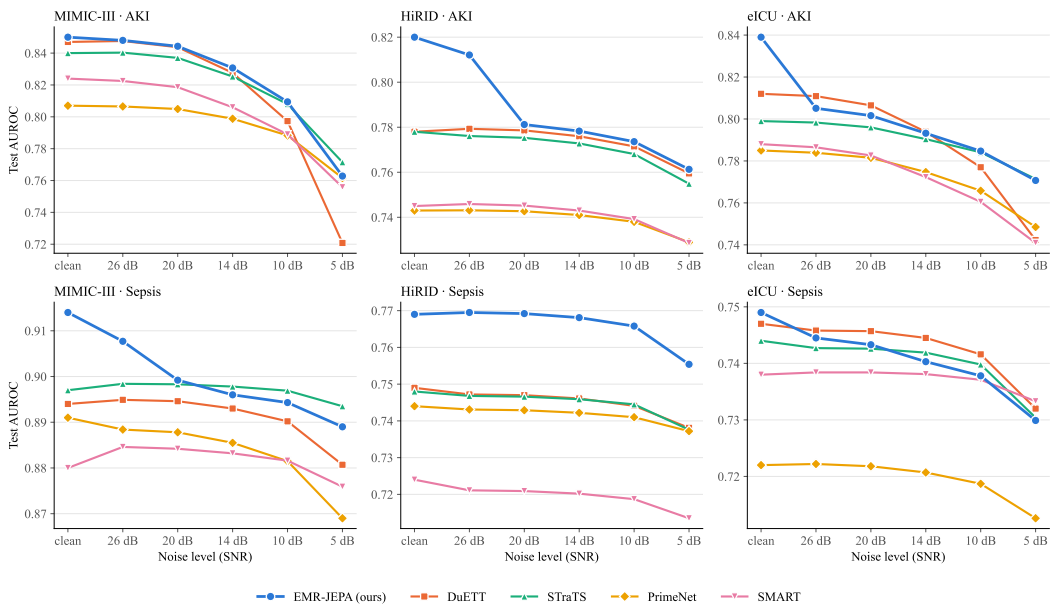



Figure B1: **Task-specific AUROC under additive value noise.** Columns show MIMIC-III, HiRID, and eICU; rows show AKI and sepsis. Numerical values are in Tables B1 and B2. The clean column matches the displayed transfer-learning means in Table 3(a). Lower SNR means stronger perturbation.

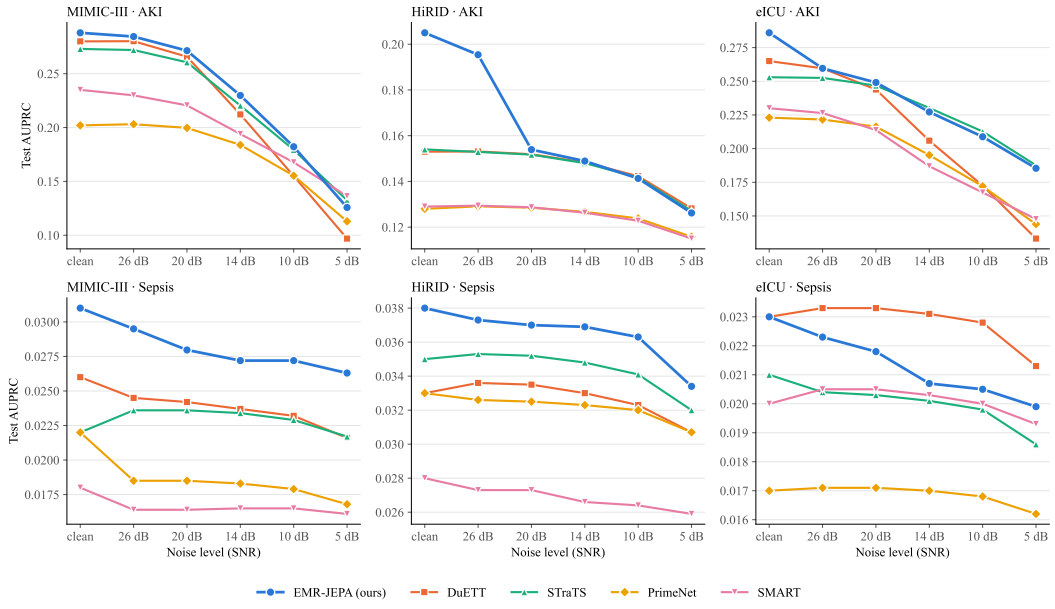



Figure B2: **Task-specific AUPRC under additive value noise.** Tasks, targets, and noise levels match Figure B1; values are in Tables B1 and B2. EMR-JEPA remains strongest on MIMIC-III and HiRID sepsis, but not on every task–target combination under corruption.

Table B1: **AKI performance under additive value noise.** Columns give the clean evaluation and five nominal SNR levels; lower SNR means stronger corruption. Values correspond to the AKI panels in Figures B1 and B2. The clean values match the displayed transfer-learning means in Table 3(a). Bold marks the largest value within each target, metric, and noise level. Seed-wise dispersion is not provided for this sweep.

| Target | Model | Metric | Clean | 26 dB | 20 dB | 14 dB | 10 dB | 5 dB |
|---|---|---|---|---|---|---|---|---|
| MIMIC-III | **EMR-JEPA** | AUROC | **0.8500** | **0.8480** | **0.8443** | **0.8307** | **0.8094** | 0.7628 |
| | | AUPRC | **0.2880** | **0.2845** | **0.2714** | **0.2298** | **0.1822** | 0.1258 |
| | DuETT | AUROC | 0.8470 | 0.8476 | 0.8437 | 0.8276 | 0.7973 | 0.7208 |
| | | AUPRC | 0.2800 | 0.2802 | 0.2658 | 0.2122 | 0.1550 | 0.0969 |
| | STraTS | AUROC | 0.8400 | 0.8403 | 0.8370 | 0.8253 | 0.8080 | **0.7715** |
| | | AUPRC | 0.2730 | 0.2720 | 0.2606 | 0.2204 | 0.1793 | 0.1325 |
| | PrimeNet | AUROC | 0.8070 | 0.8065 | 0.8049 | 0.7988 | 0.7883 | 0.7616 |
| | | AUPRC | 0.2020 | 0.2031 | 0.1997 | 0.1839 | 0.1553 | 0.1129 |
| | SMART | AUROC | 0.8240 | 0.8225 | 0.8186 | 0.8060 | 0.7890 | 0.7558 |
| | | AUPRC | 0.2350 | 0.2299 | 0.2206 | 0.1940 | 0.1677 | **0.1363** |
| HiRID | **EMR-JEPA** | AUROC | **0.8200** | **0.8121** | **0.7812** | **0.7783** | **0.7736** | **0.7613** |
| | | AUPRC | **0.2050** | **0.1954** | **0.1539** | **0.1489** | 0.1413 | 0.1262 |
| | DuETT | AUROC | 0.7780 | 0.7793 | 0.7786 | 0.7760 | 0.7715 | 0.7595 |
| | | AUPRC | 0.1530 | 0.1531 | 0.1519 | 0.1483 | **0.1423** | **0.1282** |
| | STraTS | AUROC | 0.7780 | 0.7761 | 0.7753 | 0.7728 | 0.7681 | 0.7549 |
| | | AUPRC | 0.1540 | 0.1529 | 0.1517 | 0.1480 | 0.1417 | 0.1271 |
| | PrimeNet | AUROC | 0.7430 | 0.7431 | 0.7427 | 0.7410 | 0.7380 | 0.7289 |
| | | AUPRC | 0.1280 | 0.1291 | 0.1285 | 0.1267 | 0.1238 | 0.1158 |
| | SMART | AUROC | 0.7450 | 0.7459 | 0.7452 | 0.7430 | 0.7392 | 0.7286 |
| | | AUPRC | 0.1290 | 0.1294 | 0.1287 | 0.1263 | 0.1228 | 0.1150 |
| eICU | **EMR-JEPA** | AUROC | **0.8390** | 0.8051 | 0.8016 | 0.7932 | **0.7847** | 0.7707 |
| | | AUPRC | **0.2860** | **0.2596** | **0.2491** | 0.2272 | 0.2088 | 0.1854 |
| | DuETT | AUROC | 0.8120 | **0.8109** | **0.8065** | **0.7939** | 0.7770 | 0.7423 |
| | | AUPRC | 0.2650 | **0.2596** | 0.2439 | 0.2059 | 0.1725 | 0.1332 |
| | STraTS | AUROC | 0.7990 | 0.7983 | 0.7960 | 0.7904 | 0.7841 | **0.7714** |
| | | AUPRC | 0.2530 | 0.2525 | 0.2468 | **0.2302** | **0.2127** | **0.1876** |
| | PrimeNet | AUROC | 0.7850 | 0.7839 | 0.7815 | 0.7747 | 0.7658 | 0.7485 |
| | | AUPRC | 0.2230 | 0.2216 | 0.2163 | 0.1952 | 0.1720 | 0.1439 |
| | SMART | AUROC | 0.7880 | 0.7865 | 0.7826 | 0.7723 | 0.7605 | 0.7409 |
| | | AUPRC | 0.2300 | 0.2264 | 0.2138 | 0.1869 | 0.1674 | 0.1477 |

Table B2: **Sepsis performance under additive value noise.** Columns give the clean evaluation and five nominal SNR levels; lower SNR means stronger corruption. Values correspond to the sepsis panels in Figures B1 and B2. The clean values match the displayed transfer-learning means in Table 3(a). Bold marks the largest value within each target, metric, and noise level. Seed-wise dispersion is not provided for this sweep.

| Target | Model | Metric | Clean | 26 dB | 20 dB | 14 dB | 10 dB | 5 dB |
|---|---|---|---|---|---|---|---|---|
| MIMIC-III | **EMR-JEPA** | AUROC | **0.9140** | **0.9077** | **0.8992** | 0.8960 | 0.8943 | 0.8890 |
| | | AUPRC | **0.0310** | **0.0295** | **0.02797** | **0.0272** | **0.0272** | **0.0263** |
| | DuETT | AUROC | 0.8940 | 0.8949 | 0.8946 | 0.8930 | 0.8902 | 0.8807 |
| | | AUPRC | 0.0260 | 0.0245 | 0.0242 | 0.0237 | 0.0232 | 0.0216 |
| | STraTS | AUROC | 0.8970 | 0.8984 | 0.8983 | **0.8978** | **0.8969** | **0.8935** |
| | | AUPRC | 0.0220 | 0.0236 | 0.0236 | 0.0234 | 0.0229 | 0.0217 |
| | PrimeNet | AUROC | 0.8910 | 0.8884 | 0.8878 | 0.8855 | 0.8815 | 0.8690 |
| | | AUPRC | 0.0220 | 0.0185 | 0.0185 | 0.0183 | 0.0179 | 0.0168 |
| | SMART | AUROC | 0.8800 | 0.8846 | 0.8842 | 0.8832 | 0.8816 | 0.8759 |
| | | AUPRC | 0.0180 | 0.0164 | 0.0164 | 0.0165 | 0.0165 | 0.0161 |
| HiRID | **EMR-JEPA** | AUROC | **0.7690** | **0.7695** | **0.7692** | **0.7681** | **0.7658** | **0.7554** |
| | | AUPRC | **0.0380** | **0.0373** | **0.0370** | **0.0369** | **0.0363** | **0.0334** |
| | DuETT | AUROC | 0.7490 | 0.7472 | 0.7470 | 0.7461 | 0.7441 | 0.7381 |
| | | AUPRC | 0.0330 | 0.0336 | 0.0335 | 0.0330 | 0.0323 | 0.0307 |
| | STraTS | AUROC | 0.7480 | 0.7468 | 0.7466 | 0.7459 | 0.7445 | 0.7376 |
| | | AUPRC | 0.0350 | 0.0353 | 0.0352 | 0.0348 | 0.0341 | 0.0320 |
| | PrimeNet | AUROC | 0.7440 | 0.7431 | 0.7429 | 0.7422 | 0.7410 | 0.7372 |
| | | AUPRC | 0.0330 | 0.0326 | 0.0325 | 0.0323 | 0.0320 | 0.0307 |
| | SMART | AUROC | 0.7240 | 0.7211 | 0.7209 | 0.7202 | 0.7187 | 0.7135 |
| | | AUPRC | 0.0280 | 0.0273 | 0.0273 | 0.0266 | 0.0264 | 0.0259 |
| eICU | **EMR-JEPA** | AUROC | **0.7490** | 0.7445 | 0.7433 | 0.7403 | 0.7378 | 0.7299 |
| | | AUPRC | **0.0230** | 0.0223 | 0.0218 | 0.0207 | 0.0205 | 0.0199 |
| | DuETT | AUROC | 0.7470 | **0.7458** | **0.7457** | **0.7445** | **0.7416** | 0.7320 |
| | | AUPRC | **0.0230** | **0.0233** | **0.0233** | **0.0231** | **0.0228** | **0.0213** |
| | STraTS | AUROC | 0.7440 | 0.7427 | 0.7426 | 0.7419 | 0.7398 | 0.7304 |
| | | AUPRC | 0.0210 | 0.0204 | 0.0203 | 0.0201 | 0.0198 | 0.0186 |
| | PrimeNet | AUROC | 0.7220 | 0.7222 | 0.7218 | 0.7207 | 0.7187 | 0.7126 |
| | | AUPRC | 0.0170 | 0.0171 | 0.0171 | 0.0170 | 0.0168 | 0.0162 |
| | SMART | AUROC | 0.7380 | 0.7384 | 0.7384 | 0.7381 | 0.7371 | **0.7333** |
| | | AUPRC | 0.0200 | 0.0205 | 0.0205 | 0.0203 | 0.0200 | 0.0193 |

**Separate aggregate sweep.** Figure B3 and Table B3 report the separate clean/20/10/5-dB aggregate sweep, with hospital-specific curves in Figures B4 and B5. EMR-JEPA has the highest aggregate AUROC at each level, decreasing from 0.8349 to 0.8123, but its 97.3% retention at 5 dB is not the highest. Means cover 12 target–task–seed cells per model except STraTS, which has nine. The different clean scores and incomplete STraTS coverage preclude treating this sweep as a matched-support average of the task-specific results above.

Table B3: **Separate aggregate value-noise sweep.** Values correspond to Figure B3 and average available target–task–seed evaluations. There are 12 cells per model except STraTS, which has nine. Retention is 5-dB AUROC divided by clean AUROC, expressed as a percentage. These are not averages of Tables B1 and B2.

| Model | Cells | Clean | 20 dB | 10 dB | 5 dB | Retention |
|---|---|---|---|---|---|---|
| **EMR-JEPA** | 12 | **0.8349** | **0.8332** | **0.8241** | **0.8123** | 97.3% |
| DuETT | 12 | 0.8034 | 0.8015 | 0.7869 | 0.7661 | 95.3% |
| STraTS | 9 | 0.7872 | 0.7866 | 0.7822 | 0.7742 | 98.3% |
| PrimeNet | 12 | 0.7582 | 0.7578 | 0.7543 | 0.7468 | 98.5% |
| SMART | 12 | 0.7580 | 0.7570 | 0.7512 | 0.7430 | 98.0% |

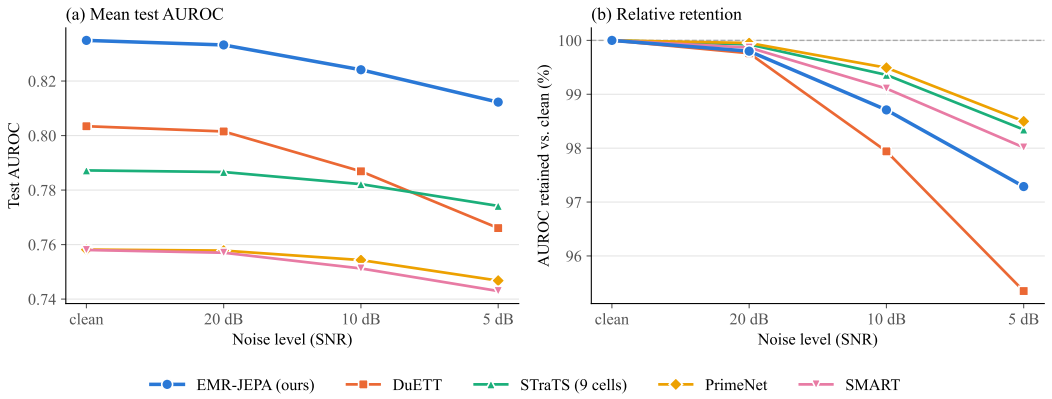



Figure B3: **Separate aggregate value-noise sweep.** Left: mean test AUROC. Right: each model's AUROC relative to its clean score. Values and support counts are in Table B3. EMR-JEPA leads absolute AUROC, not relative retention. This sweep is separate from Figures B1 and B2.

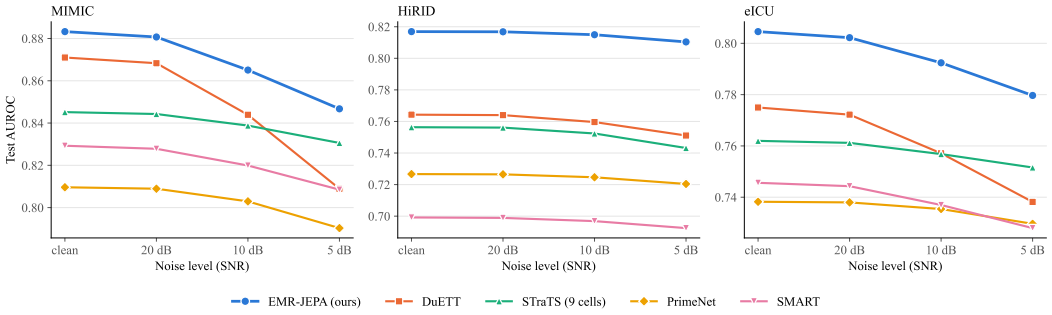



Figure B4: **Hospital-specific AUROC in the aggregate sweep.** Each panel averages available task–seed evaluations within a target hospital from the sweep in Figure B3. STraTS has incomplete support. These task-averaged curves are not the individual-task curves in Figure B1.

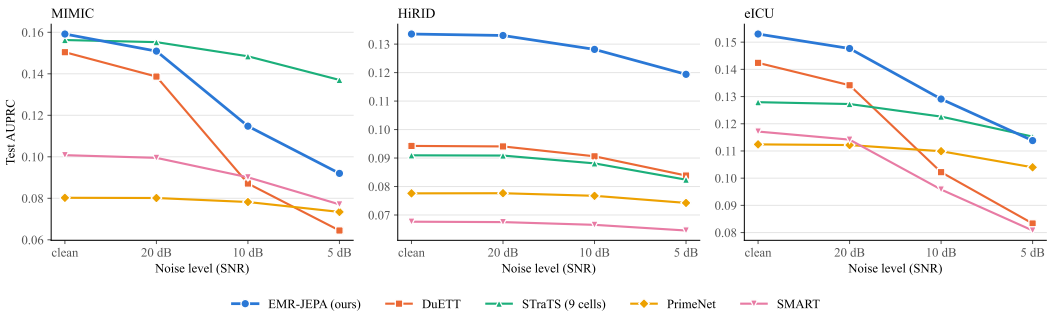



Figure B5: **Hospital-specific AUPRC in the aggregate sweep.** Settings and averaging match Figure B4. Since the averages combine tasks with different prevalence, the task-specific results in Figure B2 remain the basis for task-specific interpretation.

## C Random Initialization versus Self-Supervised Pretraining

We retain the EMR-JEPA architecture and train both the encoder and linear classifier, changing only whether training starts from random or SSL initialization. OOD evaluation trains on the two source hospitals and tests on the held-out hospital; ID evaluation averages performance on the two source test splits. Transfer trains on the target hospital, with source-domain pretraining used only for SSL initialization.

Tables C1 and C2 report higher SSL means in both metrics for all six task–fold combinations in the initialization experiment. Table C3 extends the comparison to both tasks and all three targets: SSL improves every AKI metric and every sepsis AUROC, while eICU sepsis AUPRC is slightly lower (0.023 versus 0.024). The controls support a contribution from pretraining beyond architecture alone, without implying improvement on every downstream metric.

Table C1: **Initialization control: OOD evaluation.** The encoder and classifier are trained on the two source hospitals and tested on the held-out hospital. Entries reproduce the reported means and standard deviations for random and SSL initialization. Bold marks the larger mean; it does not indicate statistical significance.

| Held-out hospital | Task | Random initialization | | SSL (Ours) | |
|---|---|---|---|---|---|
| | | AUROC | AUPRC | AUROC | AUPRC |
| MIMIC-III | AKI | $0.7765 \pm 0.001$ | $0.2031 \pm 0.001$ | $\mathbf{0.793 \pm 0.000}$ | $\mathbf{0.211 \pm 0.004}$ |
| | Sepsis | $0.6613 \pm 0.001$ | $0.0047 \pm 0.002$ | $\mathbf{0.769 \pm 0.014}$ | $\mathbf{0.008 \pm 0.001}$ |
| HiRID | AKI | $0.697 \pm 0.002$ | $0.091 \pm 0.004$ | $\mathbf{0.704 \pm 0.004}$ | $\mathbf{0.104 \pm 0.001}$ |
| | Sepsis | $0.7029 \pm 0.002$ | $0.0222 \pm 0.001$ | $\mathbf{0.756 \pm 0.005}$ | $\mathbf{0.026 \pm 0.000}$ |
| eICU | AKI | $0.693 \pm 0.018$ | $0.141 \pm 0.020$ | $\mathbf{0.7351 \pm 0.004}$ | $\mathbf{0.1644 \pm 0.002}$ |
| | Sepsis | $0.662 \pm 0.006$ | $0.011 \pm 0.000$ | $\mathbf{0.672 \pm 0.001}$ | $\mathbf{0.013 \pm 0.001}$ |

Table C2: **Initialization control: ID evaluation.** Each entry reports the mean test performance across the two source hospitals in the corresponding held-out fold. The encoder and classifier are trained jointly. Means and standard deviations are reproduced from the initialization experiment, rather than reconstructed from rounded hospital-specific entries in Table D1. Bold marks the larger mean.

| Held-out fold | Task | Random initialization | | SSL (Ours) | |
|---|---|---|---|---|---|
| | | AUROC | AUPRC | AUROC | AUPRC |
| MIMIC-III | AKI | $0.8237 \pm 0.0017$ | $0.2364 \pm 0.012$ | $\mathbf{0.8290 \pm 0.005}$ | $\mathbf{0.2455 \pm 0.004}$ |
| | Sepsis | $0.7896 \pm 0.003$ | $0.0396 \pm 0.002$ | $\mathbf{0.8045 \pm 0.005}$ | $\mathbf{0.0440 \pm 0.001}$ |
| HiRID | AKI | $0.8405 \pm 0.002$ | $0.2820 \pm 0.001$ | $\mathbf{0.8470 \pm 0.002}$ | $\mathbf{0.2950 \pm 0.000}$ |
| | Sepsis | $0.8370 \pm 0.001$ | $0.0245 \pm 0.002$ | $\mathbf{0.8525 \pm 0.002}$ | $\mathbf{0.0265 \pm 0.001}$ |
| eICU | AKI | $0.8012 \pm 0.001$ | $0.2336 \pm 0.004$ | $\mathbf{0.8305 \pm 0.001}$ | $\mathbf{0.2370 \pm 0.003}$ |
| | Sepsis | $0.8597 \pm 0.001$ | $0.0401 \pm 0.000$ | $\mathbf{0.8655 \pm 0.002}$ | $\mathbf{0.0430 \pm 0.001}$ |

Table C3: **Initialization control: target-domain transfer.** Random initialization and SSL initialization use target-domain supervised training of the encoder and classifier; SSL first pretrains on the other two hospitals. All task–target settings have reported results. Entries reproduce the supplied means and standard deviations. Bold marks the larger mean.

| Task | Target | Random initialization | | SSL (Ours) | |
|---|---|---|---|---|---|
| | | AUROC | AUPRC | AUROC | AUPRC |
| AKI | MIMIC-III | $0.845 \pm 0.001$ | $0.279 \pm 0.000$ | $\mathbf{0.850 \pm 0.000}$ | $\mathbf{0.288 \pm 0.002}$ |
| | HiRID | $0.811 \pm 0.001$ | $0.184 \pm 0.002$ | $\mathbf{0.820 \pm 0.000}$ | $\mathbf{0.205 \pm 0.004}$ |
| | eICU | $0.836 \pm 0.000$ | $0.283 \pm 0.001$ | $\mathbf{0.839 \pm 0.000}$ | $\mathbf{0.286 \pm 0.001}$ |
| Sepsis | MIMIC-III | $0.908 \pm 0.002$ | $0.024 \pm 0.001$ | $\mathbf{0.914 \pm 0.001}$ | $\mathbf{0.031 \pm 0.002}$ |
| | HiRID | $0.736 \pm 0.003$ | $0.029 \pm 0.001$ | $\mathbf{0.769 \pm 0.001}$ | $\mathbf{0.038 \pm 0.001}$ |
| | eICU | $0.729 \pm 0.004$ | $\mathbf{0.024 \pm 0.002}$ | $\mathbf{0.749 \pm 0.002}$ | $0.023 \pm 0.000$ |

## D SOURCE-DOMAIN PERFORMANCE OF MULTI-SOURCE MODELS

Table D1 evaluates each jointly trained model separately on its two source hospitals. EMR-JEPA leads both sepsis metrics and AKI AUPRC on all six source tests. The AKI AUROC exception is MIMIC-III after MIMIC-III+HiRID training, where DuETT scores $0.848$ versus $0.846$. These results characterize the retained source-domain utility of the representation; OOD generalization is measured on a different, held-out hospital.

| Task | Joint Training Domains | Evaluation Domain | Metric | SMART | DuETT | PrimeNet | STraTS | Ours |
|---|---|---|---|---|---|---|---|---|
| Sepsis | eICU + HiRID | eICU | AUROC | $0.716 \pm 0.000$ | $0.749 \pm 0.001$ | $0.702 \pm 0.000$ | $0.739 \pm 0.001$ | $\mathbf{0.786 \pm 0.003}$ |
| | | | AUPRC | $0.015 \pm 0.000$ | $0.024 \pm 0.000$ | $0.015 \pm 0.000$ | $0.022 \pm 0.000$ | $\mathbf{0.025 \pm 0.001}$ |
| | | HiRID | AUROC | $0.709 \pm 0.003$ | $0.745 \pm 0.001$ | $0.720 \pm 0.001$ | $0.747 \pm 0.002$ | $\mathbf{0.822 \pm 0.003}$ |
| | | | AUPRC | $0.026 \pm 0.000$ | $0.032 \pm 0.001$ | $0.027 \pm 0.000$ | $0.034 \pm 0.001$ | $\mathbf{0.062 \pm 0.001}$ |
| | MIMIC-III + eICU | MIMIC-III | AUROC | $0.741 \pm 0.004$ | $0.907 \pm 0.000$ | $0.834 \pm 0.004$ | $0.890 \pm 0.005$ | $\mathbf{0.922 \pm 0.002}$ |
| | | | AUPRC | $0.007 \pm 0.000$ | $0.020 \pm 0.000$ | $0.010 \pm 0.000$ | $0.023 \pm 0.001$ | $\mathbf{0.026 \pm 0.000}$ |
| | | eICU | AUROC | $0.706 \pm 0.001$ | $0.749 \pm 0.002$ | $0.701 \pm 0.001$ | $0.739 \pm 0.001$ | $\mathbf{0.782 \pm 0.000}$ |
| | | | AUPRC | $0.015 \pm 0.000$ | $0.023 \pm 0.000$ | $0.016 \pm 0.000$ | $0.021 \pm 0.000$ | $\mathbf{0.027 \pm 0.000}$ |
| | MIMIC-III + HiRID | MIMIC-III | AUROC | $0.789 \pm 0.022$ | $0.904 \pm 0.001$ | $0.867 \pm 0.000$ | $0.882 \pm 0.003$ | $\mathbf{0.918 \pm 0.001}$ |
| | | | AUPRC | $0.007 \pm 0.000$ | $0.026 \pm 0.001$ | $0.011 \pm 0.000$ | $0.021 \pm 0.001$ | $\mathbf{0.027 \pm 0.001}$ |
| | | HiRID | AUROC | $0.685 \pm 0.005$ | $0.737 \pm 0.001$ | $0.721 \pm 0.001$ | $0.742 \pm 0.001$ | $\mathbf{0.813 \pm 0.006}$ |
| | | | AUPRC | $0.025 \pm 0.001$ | $0.031 \pm 0.001$ | $0.027 \pm 0.000$ | $0.031 \pm 0.001$ | $\mathbf{0.059 \pm 0.000}$ |
| | **Average over Evaluations** | | AUROC | 0.724 | 0.799 | 0.758 | 0.790 | **0.841** |
| | | | AUPRC | 0.016 | 0.026 | 0.018 | 0.025 | **0.038** |
| AKI | eICU + HiRID | eICU | AUROC | $0.779 \pm 0.001$ | $0.809 \pm 0.000$ | $0.768 \pm 0.000$ | $0.781 \pm 0.002$ | $\mathbf{0.839 \pm 0.000}$ |
| | | | AUPRC | $0.217 \pm 0.001$ | $0.260 \pm 0.001$ | $0.206 \pm 0.000$ | $0.235 \pm 0.001$ | $\mathbf{0.285 \pm 0.001}$ |
| | | HiRID | AUROC | $0.723 \pm 0.000$ | $0.778 \pm 0.000$ | $0.718 \pm 0.000$ | $0.762 \pm 0.001$ | $\mathbf{0.819 \pm 0.000}$ |
| | | | AUPRC | $0.112 \pm 0.000$ | $0.152 \pm 0.000$ | $0.111 \pm 0.000$ | $0.146 \pm 0.001$ | $\mathbf{0.207 \pm 0.002}$ |
| | MIMIC-III + eICU | MIMIC-III | AUROC | $0.796 \pm 0.004$ | $0.846 \pm 0.001$ | $0.775 \pm 0.001$ | $0.823 \pm 0.004$ | $\mathbf{0.853 \pm 0.000}$ |
| | | | AUPRC | $0.195 \pm 0.002$ | $0.279 \pm 0.001$ | $0.180 \pm 0.000$ | $0.237 \pm 0.003$ | $\mathbf{0.300 \pm 0.002}$ |
| | | eICU | AUROC | $0.774 \pm 0.001$ | $0.809 \pm 0.000$ | $0.772 \pm 0.001$ | $0.786 \pm 0.008$ | $\mathbf{0.841 \pm 0.000}$ |
| | | | AUPRC | $0.215 \pm 0.002$ | $0.261 \pm 0.000$ | $0.210 \pm 0.000$ | $0.238 \pm 0.003$ | $\mathbf{0.288 \pm 0.000}$ |
| | MIMIC-III + HiRID | MIMIC-III | AUROC | $0.742 \pm 0.027$ | $\mathbf{0.848 \pm 0.002}$ | $0.760 \pm 0.001$ | $0.814 \pm 0.001$ | $0.846 \pm 0.000$ |
| | | | AUPRC | $0.146 \pm 0.013$ | $0.277 \pm 0.002$ | $0.146 \pm 0.001$ | $0.219 \pm 0.002$ | $\mathbf{0.284 \pm 0.005}$ |
| | | HiRID | AUROC | $0.724 \pm 0.003$ | $0.775 \pm 0.002$ | $0.733 \pm 0.000$ | $0.770 \pm 0.001$ | $\mathbf{0.815 \pm 0.003}$ |
| | | | AUPRC | $0.113 \pm 0.002$ | $0.152 \pm 0.002$ | $0.122 \pm 0.000$ | $0.149 \pm 0.001$ | $\mathbf{0.190 \pm 0.000}$ |
| | **Average over Evaluations** | | AUROC | 0.756 | 0.811 | 0.754 | 0.789 | **0.835** |
| | | | AUPRC | 0.166 | 0.230 | 0.163 | 0.204 | **0.259** |

Table D1: **Source-domain performance of multi-source models.** Models are trained on each source pair and evaluated separately on the test split of each source hospital. The table reports mean $\pm$ standard deviation over three seeds for both baselines and EMR-JEPA; supervised learning rates are selected by validation AUROC. The reported averages use the six domain-specific evaluations. These source-domain tests are distinct from evaluation on the held-out hospital.

# E Implementation and Evaluation Settings

Table E1 summarizes the documented architecture and optimization settings. EMR-JEPA fixes $(\lambda_{\text{cnt}}, \lambda_{\text{time}}) = (1, 3)$ across tasks and folds. The design studies select one SSL setting per method and fold using mean source-validation AUROC across AKI and sepsis, shared by the two tasks (Section 3). OOD fitting and selection use no target labels; transfer uses target training and validation data. Frozen-probe studies train only the classifier, whereas the main comparison and initialization controls jointly train the encoder and classifier.

Table E1: **Architecture and optimization settings.** Auxiliary weights $(1, 3)$ are shared across tasks and folds. The masking and encoding controls modify the corresponding components; frozen probes train only the downstream classifier. Seed support and selection protocols are specific to each experiment.

| Stage | Hyperparameter | Setting |
|---|---|---|
| Architecture | Model / feed-forward width | 128 / 512 |
| | Local encoder layers / heads | 2 / 4 |
| | Global encoder layers / heads | 4 / 8 |
| | Predictor layers / heads | 2 / 4 |
| | Positional encoding; dropout | Sinusoidal; 0.1 |
| | Observations per state; future queries | $p = 16$; $N = 32$ |
| Pretraining | Optimizer; weight decay | AdamW; $4 \times 10^{-3}$ |
| | Schedule; warmup fraction | Warmup–cosine; 20% |
| | Batch size; epochs | 256; 10 |
| | Gradient clipping; precision | 1.0; bf16 |
| | SSL setting selection | Source-validation AUROC, separately per fold; see Section 3 |
| EMA | Initial / final momentum | 0.996 / 1.0 (cosine schedule) |
| Losses | Latent regression | Token-reduced Smooth-L1 |
| | $(\lambda_{\text{cnt}}, \lambda_{\text{time}})$ | $(1, 3)$ for EMR-JEPA |
| Adaptation | Optimizer; batch size; epochs | AdamW; 256; 5 |
| | Schedule | Warmup–cosine (20% warmup) |
| | Checkpoint and learning-rate selection | Validation AUROC in the permitted training domain(s) |
| | Weight decay for frozen linear probes | 0 |

# F Dataset Details

We harmonize MIMIC-III, eICU-CRD, and HiRID with YAIB into 48 shared time-varying variables and static covariates (Johnson et al., 2016; Pollard et al., 2018; Hyland et al., 2020; Faltys et al., 2021; Van De Water et al., 2024).

## F.1 Cohort Construction

We follow the YAIB cohort definitions. Stays are excluded for invalid or short ($< 6$ h) length of stay, fewer than four distinct hours with measurements, a measurement gap exceeding 12 h, age below 18, or onset within the first 6 h. AKI additionally excludes baseline creatinine above 4 mg/dL (Khwaja, 2012). For eICU, sites with degenerately low task prevalence are excluded. Table F1 reports the resulting cohorts.

Table F1: Cohort attrition. Exclusion criteria follow YAIB and are applied cumulatively, so the final cohort reflects their combined effect.

| Task | Database | ICU stays | Final cohort |
|---|---|---|---|
| AKI | MIMIC-III | 61,522 | 44,216 |
| | eICU | 200,859 | 164,882 |
| | HiRID | 33,904 | 31,771 |
| Sepsis | MIMIC-III | 61,522 | 48,633 |
| | eICU | 200,859 | 123,395 |
| | HiRID | 33,904 | 29,698 |

**Sepsis labeling.** Sepsis onset follows Sepsis-3, combining suspected infection with organ dysfunction (Singer et al., 2016). Suspected infection uses antibiotic administration and culture orders for MIMIC-III and eICU, but antibiotic administration alone for HiRID because culture data are not released.

## F.2 Prediction Times and Windows

Prediction times lie on an hourly grid. A candidate time $\tau$ requires at least one observation in $[\tau-1, \tau)$ and at least 16 in the 48-hour history. The label is positive when $0 \leq \text{onset} - \tau \leq 6$ hours. Table F2 lists the windowing parameters. The future minimum is zero, so prediction times are not excluded solely for lacking future observations.

Table F2: Prediction-time and windowing parameters.

| Stage | Quantity | Value |
|---|---|---|
| Prediction time | minimum observations in last hour | 1 |
| | minimum observations in history | 16 |
| History window | window length $H$ (hours) | 48 |
| | maximum observations $M_-$ | 512 |
| | number of variables | 48 |
| Future window | minimum future observations | 0 |
| | forecast horizon $\Delta$ (hours) | 6 |
| | maximum observations $M_+$ | 512 |

## F.3 Task Cohort Statistics

Table F3 gives patient-level splits and label balance. AKI positives comprise approximately 4–6% of prediction times, whereas sepsis positives comprise approximately 0.2–1.0%, motivating AUPRC alongside AUROC. A stay contributes to these indexed splits only if it has at least one admissible prediction time; the counts can therefore be smaller than the final cohorts in Table F1.

Table F3: **Downstream task statistics.** A sample is one admissible prediction time. Its label is positive exactly when $0 \leq \text{onset} - \tau \leq 6$ hours. Negative counts equal samples minus positives. Stays differ across tasks because exclusion criteria are task-specific. Splits are disjoint at the patient level.

| Database | Split | Stays | Samples | Positive | Prev. (%) |
|---|---|---|---|---|---|
| | | | **AKI** | | |
| MIMIC-III | train | 30,991 | 1,273,437 | 55,662 | 4.37 |
| | val | 6,535 | 274,283 | 11,891 | 4.34 |
| | test | 6,531 | 266,456 | 11,874 | 4.46 |
| eICU | train | 114,620 | 3,688,363 | 230,300 | 6.24 |
| | val | 24,510 | 782,569 | 48,893 | 6.25 |
| | test | 24,653 | 808,170 | 49,084 | 6.07 |
| HiRID | train | 22,335 | 638,557 | 26,326 | 4.12 |
| | val | 4,648 | 131,351 | 5,561 | 4.23 |
| | test | 4,768 | 138,018 | 5,753 | 4.17 |
| | | | **Sepsis** | | |
| MIMIC-III | train | 34,123 | 2,005,006 | 3,324 | 0.17 |
| | val | 7,222 | 433,720 | 750 | 0.17 |
| | test | 7,156 | 420,745 | 759 | 0.18 |
| eICU | train | 85,697 | 3,940,064 | 19,255 | 0.49 |
| | val | 18,496 | 839,269 | 4,460 | 0.53 |
| | test | 18,494 | 859,534 | 4,465 | 0.52 |
| HiRID | train | 20,888 | 603,530 | 5,716 | 0.95 |
| | val | 4,326 | 123,741 | 1,264 | 1.02 |
| | test | 4,436 | 126,877 | 1,224 | 0.96 |

### F.4 Unlabeled Corpus Partitions

The task-agnostic corpus uses the intersection of the AKI and sepsis cohorts within each database, discarding labels and onset times. Patient-level train/validation/test boundaries are reused, and only source training partitions enter SSL optimization. Table F4 includes all indexed partitions: 182,698 stays and 5,609,266 prediction times in total, of which 127,987 stays and 3,927,847 prediction times belong to the training partitions. A fold uses only the training partitions of its two source hospitals.

Table F4: Unlabeled corpus partitions from the intersection of the AKI and sepsis cohorts. Only training partitions are used for SSL optimization; the totals include all three disjoint partitions. Split boundaries match Table F3.

| Database | Split | Stays | Samples |
|---|---|---|---|
| MIMIC-III | train | 29,249 | 1,139,117 |
| | val | 6,168 | 243,137 |
| | test | 6,137 | 235,816 |
| eICU | train | 78,178 | 2,316,837 |
| | val | 16,823 | 491,487 |
| | test | 16,948 | 513,561 |
| HiRID | train | 20,560 | 471,893 |
| | val | 4,265 | 96,359 |
| | test | 4,370 | 101,059 |
| MIMIC-III (total) | | 41,554 | 1,618,070 |
| eICU (total) | | 111,949 | 3,321,885 |
| HiRID (total) | | 29,195 | 669,311 |
| All | | 182,698 | 5,609,266 |

# G Encoding, Prediction, and Attention Details

## G.1 Window Construction and Patient-State Counts

With history length $H = 48$ and forecast horizon $\Delta = 6$ hours, the shifted configuration uses

$$\begin{aligned} \mathcal{O}_\tau^{\mathrm{ctx}} &= \{o_i : \tau - H \le t_i < \tau\}, \\ \mathcal{O}_\tau^{\mathrm{tgt}} &= \{o_i : \tau - (H - \Delta) \le t_i < \tau + \Delta\}. \end{aligned} \tag{6}$$

The window origins remain fixed when a stay supplies less than the full future horizon. Past and future observations are patched separately into complete groups of $p = 16$. The maximum and realized numbers of complete future states are

$$N = \lfloor M_+/p \rfloor, \qquad n = \left\lfloor \min\{|\mathcal{O}_\tau^+|, M_+\}/p \right\rfloor, \tag{7}$$

where $\mathcal{O}_\tau^+$ contains observations in $[\tau, \tau + \Delta)$ and $M_+$ is the future observation cap. The past cap is 4096 raw observations in every fold. The future cap is 2048 for the MIMIC-III- and eICU-held-out folds and 1024 for the HiRID-held-out fold, giving $N = 128$ and $N = 64$, respectively.

## G.2 Hierarchical Encoding

For observation $i$ in patch $j$, the local encoder combines relative time, scalar value, and feature embeddings. A learned token $\boldsymbol{\ell}$, shared across patches, pools the local Transformer output:

$$\begin{aligned} \mathbf{e}_{ji} &= \phi_t(t_j^\star - t_{ji}) + \phi_v(v_{ji}) + \mathbf{E}_f[f_{ji}], \\ \mathbf{h}_j &= [\mathcal{F}(\boldsymbol{\ell}, \mathbf{e}_{j1}, \ldots, \mathbf{e}_{jp})]_0 . \end{aligned} \tag{8}$$

Here $t_j^\star = \max_{o_i \in \mathcal{P}_j} t_i$ is the patch's last observation time. The scalar embedders follow STraTS (Tipirneni & Reddy, 2022). The global encoder then produces

$$\begin{aligned} \mathbf{U}^\alpha &= \mathcal{G}_\alpha\big([\mathbf{z}_{\mathrm{pol}}^\alpha, \phi_{\mathrm{sta}}^\alpha(\mathbf{x}^{\mathrm{sta}}), \mathbf{h}_{1:L_\alpha}^\alpha]; \mathbf{a}^\alpha, \mathbf{A}^\alpha\big) \\ &= [\mathbf{s}_{\mathrm{pol}}^\alpha, \mathbf{s}_{\mathrm{sta}}^\alpha, \mathbf{u}_{1:L_\alpha}^\alpha], \end{aligned} \tag{9}$$

where $\alpha \in \{\mathrm{ctx}, \mathrm{tgt}\}$ indexes the online encoder $f_\theta$ or EMA target encoder $f_\xi$. Continuous-time RoPE uses patch coordinates $a_j^{\mathrm{ctx}} = t_j^\star - (\tau - H)$ and $a_j^{\mathrm{tgt}} = t_j^\star - (\tau - (H - \Delta))$. Special tokens have no temporal coordinate. Models without policy supervision omit `POLICY`.

## G.3 Time-Conditioned Prediction and Loss Reductions

The predictor memory includes static and visible patch states but excludes direct access to the policy state. For visible historical indices $\mathcal{V}_-$, sinusoidal embeddings supplement continuous-time RoPE:

$$\begin{aligned} \mathbf{m}_j &= \mathbf{u}_j^{\mathrm{ctx}} + \mathrm{PE}(a_j^{\mathrm{ctx}} - \Delta), \\ \mathbf{M} &= [\mathbf{s}_{\mathrm{sta}}^{\mathrm{ctx}}, \{\mathbf{m}_j\}_{j \in \mathcal{V}_-}], \qquad \mathbf{q}_k = \mathbf{q} + \mathrm{PE}(a_{+,k}^{\mathrm{tgt}}), \\ \widehat{\mathbf{p}}_{1:N}^+ &= \mathcal{D}_\psi(\mathbf{q}_{1:N}, \mathbf{M}). \end{aligned} \tag{10}$$

Observed timestamps specify the $n$ valid future-query locations during pretraining; remaining slots are padded and excluded from the loss. Selected historical targets use a separate query block with timestamp and rank embeddings. Self-attention is causal within each block and blocked between the historical and future query blocks. The policy state can still influence contextualized patch and static states through global attention.

For a batch of size $B$, let $\mathcal{T}_b$ contain selected past and future targets, $\mathcal{B}_T = \{b : |\mathcal{T}_b| > 0\}$, and $\mathcal{B}_+ = \{b : n_b > 0\}$. The sample-averaged regression losses in Eq. equation 3 are

$$\begin{aligned} \mathcal{L}_{\mathrm{JEPA}} &= \frac{1}{\max(1, |\mathcal{B}_T|)} \sum_{b \in \mathcal{B}_T} \frac{\sum_{j \in \mathcal{T}_b} \sum_{r=1}^{d} \mathrm{SL1}(\widehat{p}_{bjr} - \mathrm{sg}[\bar{p}_{bjr}])}{d|\mathcal{T}_b|}, \\ \mathcal{L}_{\mathrm{time}} &= \frac{1}{\max(1, |\mathcal{B}_+|)} \sum_{b \in \mathcal{B}_+} \frac{1}{n_b} \sum_{k=1}^{n_b} (\widehat{\rho}_{bk} - \rho_{bk})^2. \end{aligned} \tag{11}$$

Count cross-entropy includes every sample, including those with no complete future state. Samples with no future states have no time loss but may still contribute historical latent targets. Denominator clamping handles batches without valid regression targets.

Starting from $\xi = \theta$, the target encoder is updated after each online optimization step:

$$\xi \leftarrow m_s \xi + (1 - m_s)\theta, \qquad m_s = 1 - \frac{1 - m_0}{2}[1 + \cos(\pi s / S)], \tag{12}$$

with $m_0 = 0.996$ and $S$ total pretraining steps.

### G.4 Attention Across Temporal Scales

Figure G1 examines all four global layers of the eICU-held-out checkpoint pretrained on MIMIC-III and HiRID. Query states are stratified by the number of raw observations within $\pm 30$ minutes of their representative timestamps; each fixed-count timestamp is the last observation in its group of 16. Attention distance is the log ratio of attention-weighted mean query–key distance to the corresponding uniform-attention mean, measured in elapsed time or intervening raw observations; static and policy tokens are excluded. Zero therefore indicates the same expected distance as uniform attention, not necessarily a uniform distribution. The policy-supervised model exhibits strongly local heads in early layers and longer-range heads in later layers, with variation across density strata. This diagnostic characterizes attention patterns without establishing a causal explanation of forecasting performance.

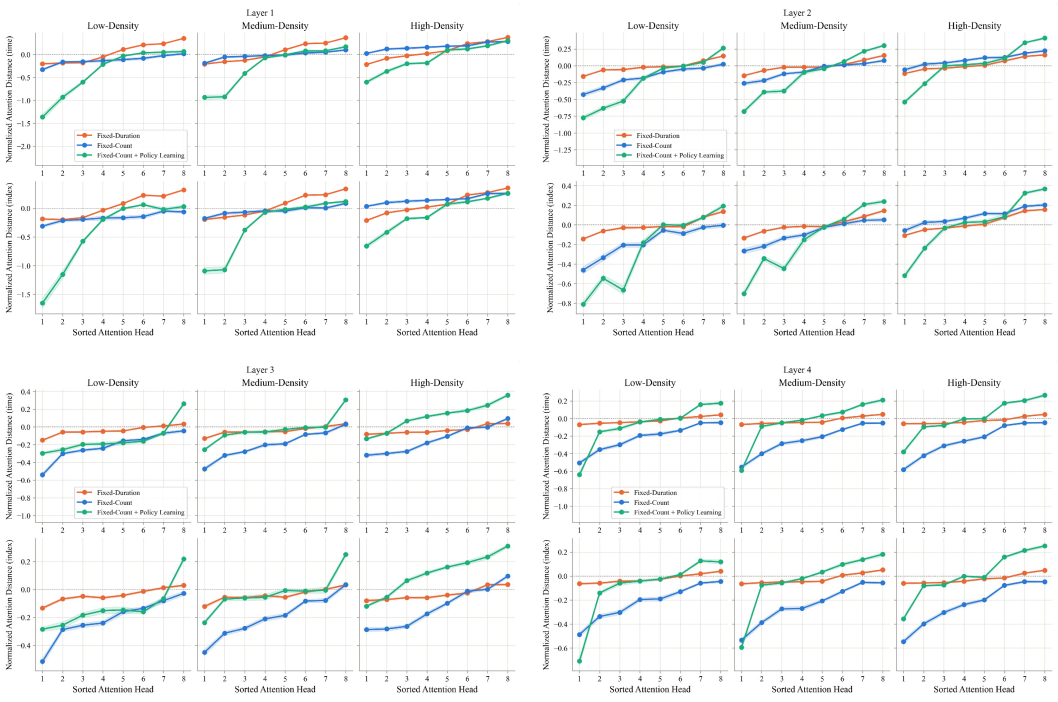



Figure G1: **Density-stratified attention spans in the eICU-held-out fold.** We compare fixed-duration encoding, fixed-count encoding, and fixed-count encoding with policy supervision across four global encoder layers. Query density is the number of raw observations within $\pm 30$ minutes of the query timestamp; each fixed-count state's timestamp is the last observation in its group of 16. Normalized attention distance is the logarithm of the ratio between the attention-weighted mean query–key distance and the corresponding mean under uniform attention. Distances are measured in elapsed time or intervening raw observations, excluding static and policy tokens. Negative and positive values indicate relatively local and longer-range attention, respectively. Heads are sorted independently for display.

# H ACQUISITION SENSITIVITY AND POLICY-FEATURE READOUTS

**Representation stability under observation thinning.** We randomly retain raw observations at keep rates 1.0, 0.5, and 0.25. For each frozen SSL checkpoint, we encode the validation split of each evaluation database with and without thinning. Cosine similarity measures alignment between corresponding representations, while linear centered kernel alignment (CKA) measures similarity of their representational geometry. Thinning is applied only during this diagnostic, never during pretraining or supervised training.

Figure H1 compares a fixed-count model pretrained without policy supervision with two readouts of a policy-supervised encoder: the full representation and its `POLICY` component. At the lowest keep rate, the full policy-supervised representation has lower cosine similarity than the separately pretrained fixed-count model across all displayed source–evaluation combinations, and the `POLICY` component changes still more. CKA shows a less uniform comparison: with MIMIC-III and eICU as sources, the policy-supervised representation retains higher CKA on HiRID. These results characterize sensitivity to the tested acquisition perturbation; they do not directly measure forecasting accuracy under thinning.

**Predictive information in the policy state.** Table H1 complements the stability diagnostic in Figure H1. We fit separate linear classifiers to two readouts of the same frozen policy-pretrained encoder in the eICU-held-out fold: the 128-dimensional policy state and the full 384-dimensional concatenation of mean patch, static, and policy states. The `POLICY`-only readout supports both AKI and sepsis forecasting, while the full readout improves AUROC and AUPRC for both tasks in ID and OOD evaluation. Together with the separately pretrained model comparison in Table 2, these results show that acquisition-sensitive information can contribute to predictive utility. This readout comparison measures accessible information within one pretrained encoder; it does not remove policy supervision from pretraining. The readouts also differ in input dimension, and global attention allows acquisition information to enter the other states.

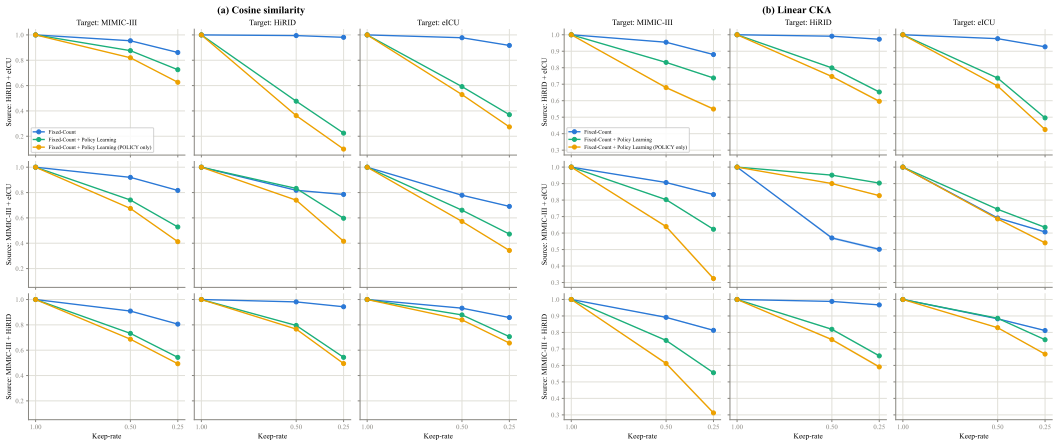



Figure H1: **Representation stability under observation thinning.** Rows identify the two pretraining source databases and columns identify the evaluation database. Frozen SSL checkpoints encode validation records before and after random thinning. Panels report cosine similarity and linear CKA relative to the unthinned representations. The keep rate denotes the fraction of raw observations retained. Fixed-Count uses a model pretrained without policy supervision; the other curves use either the full representation or only the `POLICY` state of a policy-supervised encoder. Lower similarity indicates greater representational change under thinning.

Table H1: **Predictive information in the POLICY state.** Frozen linear probes use the same policy-pretrained encoder from the eICU-held-out fold, with MIMIC-III and HiRID as sources. POLICY-only uses the 128-dimensional policy state. Full uses the 384-dimensional concatenation of mean patch, static, and policy states. OOD evaluates eICU; ID summarizes the two sources. The supervised learning rate is denoted by $\eta$. Bold indicates the better readout within each task and metric.

| Task | Readout | $\eta$ | OOD | | ID | |
|---|---|---|---|---|---|---|
| | | | AUROC | AUPRC | AUROC | AUPRC |
| AKI | POLICY-only | 0.01 | 0.6276 | 0.1007 | 0.7514 | 0.1359 |
| | Full | 0.003 | **0.6406** | **0.1044** | **0.7682** | **0.1523** |
| Sepsis | POLICY-only | 0.001 | 0.6758 | 0.0103 | 0.8357 | 0.0305 |
| | Full | 0.003 | **0.6770** | **0.0114** | **0.8512** | **0.0355** |