PAID: PATIENT-IDENTITY CONTRASTIVE LEARNING FROM RECURRENT ADMISSIONS IN MULTIMODAL EHRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable representation learning from multimodal electronic health records (EHRs) is difficult because outcome labels are task-specific and each admission provides an incomplete view of patient state. Recurrent admissions offer outcome-independent supervision: they share patient-level context while differing in acute condition and modality availability. We propose Patient-Identity Contrastive Learning (PAID), which uses this relation in an asymmetric dual-space Transformer. A patient token (PAT) is aligned across admissions, while an admission token (ADM) retains a separate route for episode-specific evidence. Availability-masked attention, cross-modal reconstruction, and a structured matching score support learning from incomplete observations. On patient-disjoint MIMIC-IV and eICU cohorts, PAID achieves the highest AUROC on all 12 endpoints, with maximum gains of 0.0135 and 0.0247, respectively. Its prediction systems improve all 21 evaluated few-shot settings, and PAT achieves 51.4% Recall@5 for same-patient retrieval. These results support recurrent admissions as a useful source of supervision for reusable multimodal representations with task-dependent patient and admission roles.

1 INTRODUCTION

Electronic health records (EHRs) provide a longitudinal account of chronic conditions, acute episodes, clinical decisions, and care across admissions. Representation learning turns this heterogeneous record into reusable features. It allows abundant unlabeled data to serve multiple prediction tasks, reduces task-specific feature engineering, and improves learning when annotated cohorts are small Rajkomar et al. (2018); Li et al. (2020); Rasmy et al. (2021). This role is especially important for multimodal EHRs because structured variables, physiological signals, clinical text, imaging, and ECG describe different aspects of patient state and are rarely collected with uniform coverage. Most representation models nevertheless treat each admission as an independent sample. They can recognize patterns within an admission while discarding the clinically informative fact that the same person may return with a persistent health background. We study how recurrent admissions can supervise patient-stable representations without suppressing the acute evidence that distinguishes one admission from another.

Existing EHR methods obtain supervision from several sources. Outcome-supervised systems optimize clinical endpoints but require labeled cohorts and often produce task-specific representations Rajkomar et al. (2018). BEHRT and Med-BERT learn within-record structure through masked modeling Li et al. (2020); Rasmy et al. (2021). CLMBR predicts future events Steinberg et al. (2021), while SimCLR and SCARF construct views through synthetic corruption Chen et al. (2020); Bahri et al. (2022). These objectives do not explicitly distinguish patient context from episode-specific evidence across two genuine admissions. Longitudinal history improves clinical prediction Miotto et al. (2016); Choi et al. (2016); Steinberg et al. (2021), but concatenating all prior observations can mix chronic background with acute state and modality availability. Patient identity has already proved useful for contrastive learning of ECGs and medical time series Diamant et al. (2022); Wang et al. (2023). It is available without outcome labels and links admissions that share persistent clinical context while differing in their immediate causes, treatments, and observations. A useful representation must capture this shared context without removing changes that distinguish the admissions.

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.

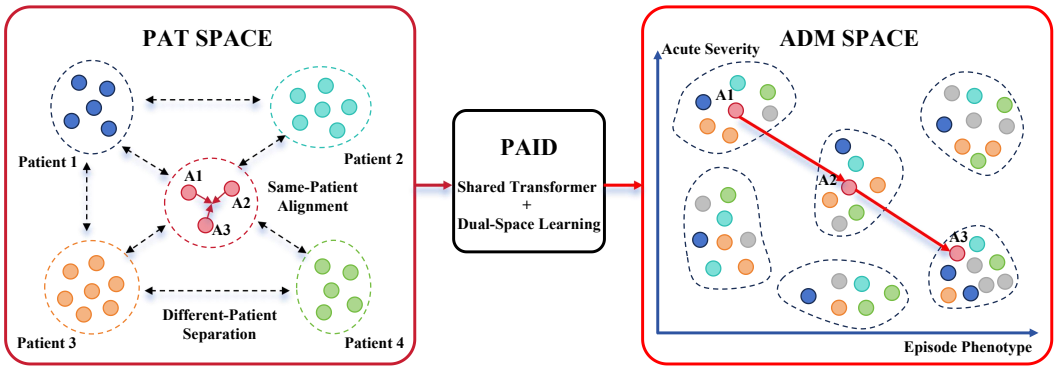

Figure 1: Supervision from recurrent admissions. A_1, A_2, and A_3 denote three admissions of one patient. Their PAT representations are aligned while representations from different patients remain separated. In ADM space, the same admissions may vary with acute severity and episode phenotype. Positions are schematic.

Recurrent admissions can be viewed as imperfect observations of a shared and slowly varying patient state. Their differences are equally important because acute disease, intervention, and available modalities may change sharply. The central challenge is to retain patient context across admissions while preserving episode-specific differences, even when the two admissions expose different subsets of the underlying clinical state.

We introduce *Patient-Identity Contrastive Learning* (PAID). Figure 1 illustrates its central learning relation. PAID maps each admission into a patient token (PAT) and an admission token (ADM). An asymmetric objective makes PAT comparable across recurrent admissions while leaving ADM to represent evidence that distinguishes the current admission. The complete architecture in Figure 2 extends this separation to incomplete multimodal records. The resulting representation uses patient identity as outcome-independent supervision without making the entire admission invariant.

There are four contributions in this paper.

- We formulate patient-identity contrast for incomplete multimodal admissions, using recurrent episodes as natural positive views while preserving a separate representation for their clinical differences.
- We introduce an asymmetrically trained dual-space encoder. Identity alignment acts on PAT, while adversarial identity suppression acts on ADM. This asymmetric training promotes cross-admission consistency in PAT and retains a separate route for admission-specific evidence in ADM.
- We develop availability-masked attention, cross-modal reconstruction, and a structured matching score to learn from admissions with different modality coverage.
- We evaluate PAID on MIMIC-IV and eICU against supervised and self-supervised systems, with mean AUROC 0.8379 on MIMIC and gains up to 0.0247 over the strongest eICU comparator. Retrieval, few-shot, module ablation, and specialization analyses examine the learned representation.

2 Related Work

Clinical representation learning. Sequence models and clinical foundation models learn representations from codes, events, or notes through objectives such as next-event prediction and masked modeling Rajkomar et al. (2018); Li et al. (2020); Rasmy et al. (2021); Steinberg et al. (2021). Recent work extends this direction through time-to-event pretraining in MOTOR Steinberg et al. (2024), generative patient timelines in Foresight Kraljevic et al. (2024), long-context EHR modeling Wornow et al. (2025), and cross-site adaptation of a shared EHR foundation model Guo et al. (2024). Generative Deep Patient combines structured time series and clinical text within a multimodal generator Sivarajkumar et al. (2026). These methods learn from event sequences or complete

timelines. PAID instead contrasts distinct admissions of an identified patient and assigns different identity constraints to patient and admission tokens.

Contrastive learning. SimCLR and MoCo form positive pairs through augmentation and use other samples as negatives Chen et al. (2020); He et al. (2020b). Supervised contrastive learning extends this construction to samples with a shared class label Khosla et al. (2020). SCARF applies contrastive learning to tabular data through feature corruption Bahri et al. (2022). PCLR contrasts ECGs from the same patient Diamant et al. (2022), and COMET includes patient-level relations in a hierarchy of time-series contrasts Wang et al. (2023). PAID builds on patient-based supervision through asymmetric PAT/ADM training for recurrent multimodal admissions with incomplete modality coverage.

Multimodal and incomplete clinical data. Multimodal clinical systems must accommodate irregular sampling and modalities that may be absent by design rather than missing at random. Masked autoencoders provide a general way to learn from partially observed inputs He et al. (2022), while representation regularizers such as VICReg prevent degenerate contrastive solutions Bardes et al. (2022). MUSE learns multimodal patient representations with missing modalities and labels using a patient–modality graph Wu et al. (2024). PAID instead pairs observed admissions and adapts reconstruction and regularization to make same-patient admissions comparable under incomplete modality coverage while retaining an explicit patient-level objective.

3 METHOD

3.1 PROBLEM FORMULATION

Let i index an admission and $p_i \in \mathcal{P}$ identify its patient. The admission contains up to M modalities $\mathbf{X}_i = \{\mathbf{x}_i^{(m)}\}_{m=1}^{M}$ with availability $\mathbf{a}_i \in \{0,1\}^M$. Patient identity defines

$$y_{ik} = \begin{cases} +1, & p_i = p_k,\ i \neq k, \\ -1, & p_i \neq p_k. \end{cases} \tag{1}$$

Positive pairs are distinct clinical episodes that may differ in acute condition, treatment, and coverage. Pair construction requires no downstream outcome label. We seek an encoder $E_\theta(\mathbf{X}_i, \mathbf{a}_i) = (\mathbf{h}_i^{\mathrm{pat}}, \mathbf{h}_i^{\mathrm{adm}}, \sigma_i)$. PAT should retain information reproducible across admissions, while ADM preserves a route for episode-specific evidence. The scalar σ_i adjusts identity matching for token dispersion, context, and missingness. The two tokens provide complementary readouts for downstream clinical tasks.

3.2 PAID FRAMEWORK

Figure 2 organizes PAID into Longitudinal Views, the PAID Encoder, and Representation Spaces. Longitudinal Views pair recurrent admissions and use other patients as negatives. In the PAID Encoder, modality tokenizers and type embeddings form the inputs labeled Embedding. Availability and reconstruction masks control attention in the shared Transformer. Its PAT and ADM tokens produce the two representation spaces. Identity alignment acts on PAT and adversarial suppression on ADM, supported by reconstruction and the matching score under incomplete coverage. The framework requires patient linkage and availability masks without constraining the upstream modality encoders.

3.3 TRANSFORMER-BASED DUAL-SPACE ENCODER

A pre-normalized Transformer processes modality tokens. Each modality-specific tokenizer $T_m : \mathcal{X}_m \to \mathbb{R}^d$ maps its input to a common width, followed by a learned type embedding:

$$\mathbf{u}_i^{(m)} = T_m(\mathbf{x}_i^{(m)}) + \mathbf{e}_m. \tag{2}$$

For reconstruction mask $\mathbf{r}_i$, the visible-modality vector is $\widetilde{\mathbf{a}}_i = \mathbf{a}_i \odot (\mathbf{1} - \mathbf{r}_i)$. At inference $\mathbf{r}_i = \mathbf{0}$. The Transformer input contains two learned aggregation tokens followed by the modality tokens

$$\mathbf{S}_i = [\mathbf{c}_{\mathrm{pat}}, \mathbf{c}_{\mathrm{adm}}, \mathbf{u}_i^{(1)}, \ldots, \mathbf{u}_i^{(M)}].$$

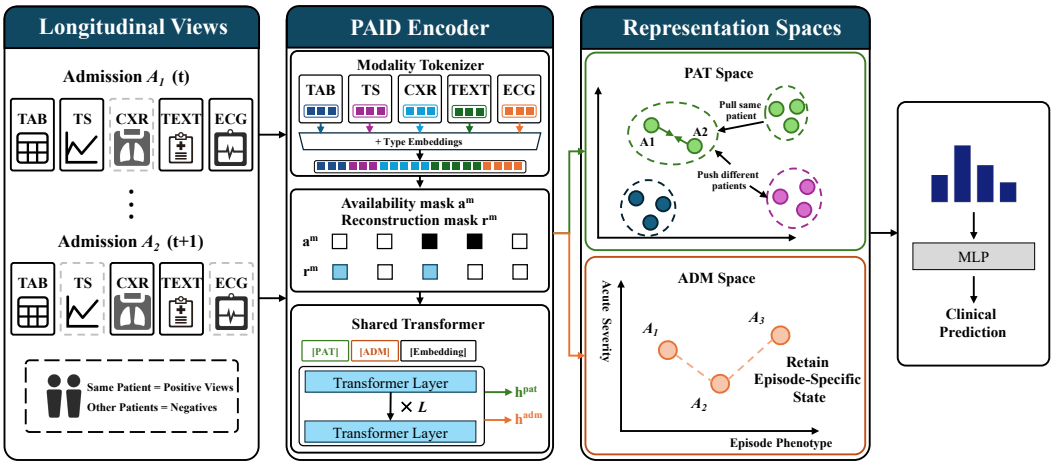

Figure 2: PAID architecture for two recurrent admissions. TAB, TS, CXR, TEXT, and ECG denote tabular variables, time series, chest radiographs, clinical text, and electrocardiograms. A shared masked Transformer produces PAT for identity alignment and ADM for episode-specific evidence. Training-only heads and the score readout are omitted.

Let $\mathbf{H}_i^0 = \mathbf{S}_i$. Each of the L Transformer blocks applies masked multi-head self-attention and a feed-forward network with pre-normalization and residual connections.

$$\begin{aligned}\bar{\mathbf{H}}_i^\ell &= \mathbf{H}_i^{\ell-1} + \mathrm{MHA}(\mathrm{LN}(\mathbf{H}_i^{\ell-1}), \mathbf{A}_i),\\ \mathbf{H}_i^\ell &= \bar{\mathbf{H}}_i^\ell + \mathrm{FFN}(\mathrm{LN}(\bar{\mathbf{H}}_i^\ell)).\end{aligned} \tag{3}$$

Here $(\mathbf{A}_i)_{qk} = -\infty$ if key k denotes a modality m with $\widetilde{a}_i^{(m)} = 0$, and is zero otherwise. Both aggregation tokens remain visible, so availability is integrated into every Transformer layer. The final outputs are $\mathbf{h}_i^{\mathrm{pat}} = \mathbf{H}_{i,1}^L$ and $\mathbf{h}_i^{\mathrm{adm}} = \mathbf{H}_{i,2}^L$.

Aligning a single representation can suppress acute changes needed by downstream tasks. PAT therefore receives identity attraction, while ADM remains outside that objective. To discourage identity recovery from ADM, we train an in-batch matcher through gradient reversal Ganin et al. (2016). Let $\mathbf{q}_i^e = \mathrm{norm}(g_{\mathrm{adm}}(\mathbf{h}_i^{\mathrm{adm},e}))$ and $\mathbf{q}_k^l$ denote projections for paired earlier and later admissions. With $C_{ik} = (\mathbf{q}_i^e)^\top \mathbf{q}_k^l / \tau_{\mathrm{adm}}$, the matcher minimizes

$$\mathcal{L}_{\mathrm{adv}} = \tfrac{1}{2}\left[\mathrm{CE}(\mathbf{C}, \mathbf{I}) + \mathrm{CE}(\mathbf{C}^\top, \mathbf{I})\right], \tag{4}$$

where the diagonal identifies the true pairing. We use $\tau_{\mathrm{adm}} = \max(|\theta|, 0.01)$ with learned θ initialized to 0.07. Gradient reversal multiplies the encoder gradient by $-\alpha_e$, where $\alpha_e = 2\min(1, e/10)$ at epoch e, without changing the matcher update. This encourages identity consistency in PAT and less recoverable identity in ADM. The modality tokenizers, Transformer, and aggregation tokens form the backbone. The projection, adversarial, reconstruction, semantic, and score heads specify its training and readout.

3.4 Uncertainty-Aware Patient-Identity Contrast

Paired admissions can differ in information quality and modality coverage. We adjust the matching logit with a nonnegative score from three sources. Let $\mathbf{v}_i^{(m)} = T_m(\mathbf{x}_i^{(m)})$ be the tokenizer output before adding a type embedding, $n_i = \sum_m \widetilde{a}_i^{(m)}$, and $\bar{\mathbf{v}}_i = n_i^{-1} \sum_m \widetilde{a}_i^{(m)} \mathbf{v}_i^{(m)}$. The first term measures dispersion across visible token coordinates:

$$\sigma_i^{\mathrm{disp}} = \frac{1}{d}\sum_{r=1}^{d}\left[\frac{1}{n_i^2}\sum_m \widetilde{a}_i^{(m)}(v_{ir}^{(m)} - \bar{v}_{ir})^2 + \epsilon\right]^{1/2}. \tag{5}$$

The remaining components are learned from the patient context and missingness pattern. Define $d_i^{(m)} = \mathrm{softplus}(b_m) + \mathrm{softplus}(f_{\mathrm{miss}}(\mathbf{h}_i^{\mathrm{pat}})_m)$.

$$\begin{aligned} \sigma_i^{\mathrm{ctx}} &= \mathrm{softplus}(f_{\mathrm{ctx}}(\mathbf{h}_i^{\mathrm{pat}})), \\ \sigma_i^{\mathrm{miss}} &= \sum_m (1 - \widetilde{a}_i^{(m)})\, d_i^{(m)}, \\ \sigma_i^2 &= (\sigma_i^{\mathrm{disp}})^2 + (\sigma_i^{\mathrm{ctx}})^2 + (\sigma_i^{\mathrm{miss}})^2 + \epsilon. \end{aligned} \tag{6}$$

Observed modalities contribute zero to the missingness term. Together, these components adapt identity matching to the available evidence. Appendix A.4 derives the score's properties and gradients.

We project the PAT output as $\mathbf{z}_i = g_{\mathrm{pat}}(\mathbf{h}_i^{\mathrm{pat}})/\|g_{\mathrm{pat}}(\mathbf{h}_i^{\mathrm{pat}})\|$. For an earlier admission i and a later admission k, the identity score is

$$s_{ik} = \exp(t)\left[\mathbf{z}_i^\top \mathbf{z}_k - \gamma(\sigma_i + \sigma_k)\right] + b. \tag{7}$$

We use $\gamma = 0.1$. The scale $\exp(t)$ and bias b are learned. The score affects representation learning: increasing σ lowers the logit, raising the positive-pair loss and lowering the negative-pair loss. With Eq. 1, a class-balanced pairwise sigmoid objective inspired by sigmoid contrastive learning Zhai et al. (2023) gives

$$\mathcal{L}_{\mathrm{id}} = \frac{1}{Z} \sum_{i,k} w_{y_{ik}} \log(1 + \exp(-y_{ik} s_{ik}))\,. \tag{8}$$

Here $w_+ = N_-/N_+$, $w_- = 1$, and $Z = w_+ N_+ + w_- N_-$ is the sum of pair weights. Unlike row-normalized InfoNCE He et al. (2020a), it classifies each pair directly.

Each batch draws B distinct patients and an ordered pair of admissions per patient. The B matched cross-view pairs are positive, and the remaining $B^2 - B$ pairs are negatives. Patients with more than two admissions expose different ordered pairs across epochs.

3.5 Cross-Modal Learning from Incomplete Admissions

Cross-modal learning gives the encoder a direct content-learning signal alongside patient alignment. It trains visible inputs to predict withheld clinical information when paired admissions expose different modality subsets. For each admission, we sample $\mathcal{R}_i \subseteq \{m \mid a_i^{(m)} = 1\}$ from the observed set. Attention excludes hidden modalities as keys and values. Before decoding, their residual token outputs are replaced by learned mask tokens, giving $\widetilde{\mathbf{H}}_i^L$. A lightweight decoder predicts $\hat{\mathbf{x}}_i^{(m)} = D_m(\widetilde{\mathbf{H}}_i^L)$ from this visible context (Appendix A.5). Let $\mathcal{B}_m = \{i : r_i^{(m)} = 1\}$ and $\mathcal{M}_{\mathcal{R}} = \{m \mid |\mathcal{B}_m| > 0\}$. The reconstruction loss is

$$\begin{aligned} \ell_m &= \frac{1}{|\mathcal{B}_m| d_m} \sum_{i \in \mathcal{B}_m} \left\| \mathrm{LN}(\hat{\mathbf{x}}_i^{(m)}) - \mathrm{LN}(\mathbf{x}_i^{(m)}) \right\|_2^2, \\ \mathcal{L}_{\mathrm{mae}} &= \frac{1}{|\mathcal{M}_{\mathcal{R}}|} \sum_{m \in \mathcal{M}_{\mathcal{R}}} \ell_m. \end{aligned} \tag{9}$$

Coordinate normalization limits differences in feature scale. Masking retains visible anchor inputs (Appendix A.5). Reconstruction learns predictive relations across observed subsets. Decoder outputs are discarded after training and are not used to impute downstream inputs.

When an automatically derived semantic target is available, a small head adds masked binary cross-entropy $\mathcal{L}_{\mathrm{sem}}$. The term is zero when no target is observed. MIMIC uses eight CheXpert findings Irvin et al. (2019): early PAT predicts the later admission's findings, falling back to the earlier target if unavailable. These auxiliary labels differ from the evaluated outcomes and provide additional semantic supervision.

3.6 Unified Learning Objective and Transfer

PAID groups its objectives by modeling role. The identity group defines what should be shared across admissions. The modality group makes different observed views comparable, and the regu-

larization group stabilizes the learned representation.

$$\begin{aligned}\mathcal{L}_{\text{identity}} &= \mathcal{L}_{\text{id}} + 0.2\mathcal{L}_{\text{adv}},\\ \mathcal{L}_{\text{modality}} &= 0.5\mathcal{L}_{\text{mae}} + 0.2\mathcal{L}_{\text{sem}},\\ \mathcal{L}_{\text{reg}} &= 0.3\mathcal{L}_{\text{vic}} + 0.1\mathcal{L}_{\text{temp}} + 0.05\mathcal{L}_{\text{prior}},\\ \mathcal{L}_{\text{I}} &= \mathcal{L}_{\text{identity}} + \mathcal{L}_{\text{modality}} + \mathcal{L}_{\text{reg}}.\end{aligned} \tag{10}$$

Appendix A.6 specifies the auxiliary losses in Eq. 10. $\mathcal{L}_{\text{vic}}$ is the variance and covariance anti-collapse regularizer Bardes et al. (2022). $\mathcal{L}_{\text{temp}} = \mathbb{E}_{i \prec k}[\sigma_k - \sigma_i]_+$ provides a soft ordering constraint on the matching score. $\mathcal{L}_{\text{prior}}$ aligns PAT with a low-rank, training-only patient prototype through symmetric stop-gradient alignment.

Stage I jointly trains the backbone and all objective groups. Each step encodes the paired admissions with shared parameters and independent masks, constructs the $B \times B$ relation matrix, and optimizes Eq. 10. An optional Stage II freezes the representation backbone and fits a fresh score head using temporal and missingness-ordering constraints. Appendix A.6 gives the objective under fixed PAT and ADM representations.

The prototype, decoder, semantic head, and adversarial matcher are removed at inference. Downstream readouts use PAT, ADM, or their concatenation with optional σ, with a frozen or fine-tuned encoder as specified by the experiment. PAT retrieval tests cross-admission consistency. The shared-encoder readout comparison examines specialization and the value of combining tokens.

4 Experiments

4.1 Datasets and Preprocessing

MIMIC-IV contains deidentified hospital and emergency department records Johnson et al. (2023). Of 79,535 patients with recurrent admissions, 73,569 are used for representation learning and 5,966 for testing. The 101,198 single-admission patients are divided into 93,609 training and 7,589 test patients. All splits are patient-disjoint. Appendix A.1 gives feature dimensions and admission counts. Endpoint definitions and input-window details are documented in Appendix A.2.

Each MIMIC admission is represented by five clinical modalities. Tabular and time-series features are standardized, CXR and notes are encoded with RAD-DINO Pérez-García et al. (2025) and Clinical ModernBERT Lee et al. (2025), and ECG features are derived with NeuroKit2 Makowski et al. (2021). All comparison methods use the same feature-extraction pipeline, including XGBoost, which concatenates these pre-extracted modality features. Appendix A.3 describes method-specific adapters.

eICU is a multi-center critical-care database with 200,859 stays from 139,367 patients Pollard et al. (2018). It provides tabular, time-series, and text data but no CXR or ECG. The 38,483 patients with recurrent admissions are divided into 35,596 training and 2,887 test patients. The 100,884 single-stay patients are divided into 93,318 training and 7,566 test patients. The patient-disjoint test set contains 15,026 stays, of which 3,718 are eligible for the subsequent-stay endpoint.

4.2 Evaluation and Metrics

MIMIC covers in-hospital mortality, length of stay above 3 or 7 days, ICU utilization during the admission, readmission within 30 days after discharge, AKI, and sepsis. eICU similarly covers mortality, stays longer than 3 or 7 days, and sepsis. Table 2 also includes an exploratory subsequent-stay endpoint. The clinical comparisons evaluate retrospective classification. Appendix A.2 defines the evaluation scope.

The evaluation addresses four questions. **Q1 (Identity consistency).** Does PAT retrieve a different admission of the same patient? **Q2 (Clinical utility).** How does the complete PAID prediction system compare with supervised and self-supervised alternatives? **Q3 (Label efficiency).** Does the representation remain useful with 1%, 5%, or 10% of outcome labels? **Q4 (Representation specialization).** Do PAT and ADM serve distinct patient- and admission-level roles?

AUROC is the primary metric for binary clinical outcomes. Mean AUROC is used for model-capacity and ablation summaries. Patient retrieval is measured by Recall@1, Recall@5, and Re-

Table 1: MIMIC AUROC on patient-disjoint test cohorts. Bold marks the best point estimate.

Method	Mortality	LOS >3d	LOS >7d	ICU util.	Readm.	AKI	Sepsis
PAID	**0.9179**	**0.8480**	**0.8409**	**0.9147**	**0.6202**	**0.8319**	**0.8918**
XGBoost	0.9011	0.8473	0.8343	0.9073	0.5751	0.8249	0.8743
MLP	0.9026	0.7852	0.7892	0.8822	0.5842	0.8041	0.8655
HAIM	0.9026	0.8229	0.7891	0.8871	0.5731	0.8044	0.8623
MUSE	0.8713	0.8268	0.8325	0.9015	0.6172	0.8308	0.8741
AutoEncoder	0.8928	0.7612	0.7784	0.8649	0.5863	0.7952	0.8624
SimCLR	0.9126	0.8200	0.8135	0.8922	0.6067	0.8209	0.8882
SCARF	0.8313	0.7791	0.7401	0.7827	0.5696	0.7399	0.8165
Med-BERT	0.8926	0.7550	0.7566	0.8588	0.5903	0.7760	0.8391
CLMBR	0.8380	0.7238	0.7387	0.8184	0.5504	0.7624	0.8063
SupCon	0.7821	0.7333	0.7490	0.8322	0.5727	0.7568	0.8075

call@10. The specialization analysis also reports the mean AUROC over the six non-readmission outcomes.

4.3 Training Details and Baselines

PAID uses a six-layer Transformer with six attention heads and width 384. Pretraining lasts 100 epochs with AdamW, an initial learning rate of 5×10^{-4}, weight decay of 0.05, batch size 256, and gradient clipping at 1.0. Ten warm-up epochs are followed by cosine learning-rate decay.

The comparisons include XGBoost Chen & Guestrin (2016), an MLP Rumelhart et al. (1986), and the multimodal HAIM framework Soenksen et al. (2022), alongside a MUSE adaptation Wu et al. (2024). Representation-learning comparisons include an AutoEncoder Hinton & Salakhutdinov (2006), SimCLR Chen et al. (2020), SCARF Bahri et al. (2022), Med-BERT Rasmy et al. (2021), CLMBR Steinberg et al. (2021), and supervised contrastive learning Khosla et al. (2020). Each representation-learning method retains its pretraining objective and uses the common feature pipeline. Tables 1–2 compare complete prediction systems. Their labeled training pools and readout selection differ, as detailed in Appendix A.3.

4.4 Comparison with Existing Methods

MIMIC. Table 1 reports seven MIMIC endpoints. PAID ranks first on each endpoint. Its gain over the strongest non-PAID result ranges from 0.0007 for length of stay above 3 days to 0.0135 for readmission, with a mean gain of 0.0063. SimCLR is the strongest comparator for mortality, readmission, and sepsis. XGBoost is strongest among the remaining systems for both length-of-stay endpoints, ICU utilization, and AKI. PAID surpasses both families, so its improvement is not confined to either nonlinear tabular fitting or synthetic-view contrast. The largest margin occurs for readmission, where information shared across admissions is directly relevant. Gains on mortality, sepsis, and the other acute endpoints show that the combined representation also retains acute evidence.

4.5 Ablation Study

Table 3 groups the frozen Small-backbone ablations by the three roles in Eq. 10. Each row removes one component while retaining the other objectives. The complete model achieves the highest mean AUROC. Removing dual space or adversarial separation reduces it by 0.0568 or 0.0274, supporting the asymmetric representation design. Reconstruction and semantic prediction provide substantial content supervision, while the three regularizers consistently improve the complete system.

4.6 Representation and Label-Efficiency Analyses

Cross-admission identity (Q1). PAT retrieval uses 25,117 admissions from 5,966 test patients, ranking all other admissions by cosine similarity. Recall@1/5/10 reaches 28.4%, 51.4%, and 61.9%,

eICU. PAID leads the four clinical classification endpoints, improving over XGBoost by 0.0018 for mortality, 0.0151 for LOS >3 days, 0.0247 for LOS >7 days, and 0.0109 for sepsis. The largest margin is on prolonged stay, where the fused representation improves over both feature-based and representation-learning comparators.

Table 2: Complete eICU AUROC comparison. $n = 15{,}026$, except Next ICU ($n = 3{,}718$). †Exploratory subsequent-stay endpoint (Appendix A.10).

Method	Mort.	LOS3	LOS7	Sepsis	Next ICU†
PAID	**0.8751**	**0.8002**	**0.7567**	**0.9963**	**0.8181**
XGBoost	0.8733	0.7851	0.7320	0.9854	0.8046
AutoEncoder	0.8145	0.7317	0.6969	0.9702	0.6949
SimCLR	0.7860	0.7087	0.6668	0.8402	0.6692
SCARF	0.7496	0.7009	0.6719	0.8301	0.7323
Med-BERT	0.7476	0.6809	0.6506	0.8412	0.6314
CLMBR	0.7334	0.6859	0.6568	0.8489	0.6442
SupCon	0.7191	0.6770	0.6415	0.8324	0.6506

eICU contains no CXR or ECG, so this result extends PAID's utility to a multicenter setting with only tabular, time-series, and text inputs. Its lead across these endpoints supports the flexibility of the multimodal encoder under a systematically different modality inventory.

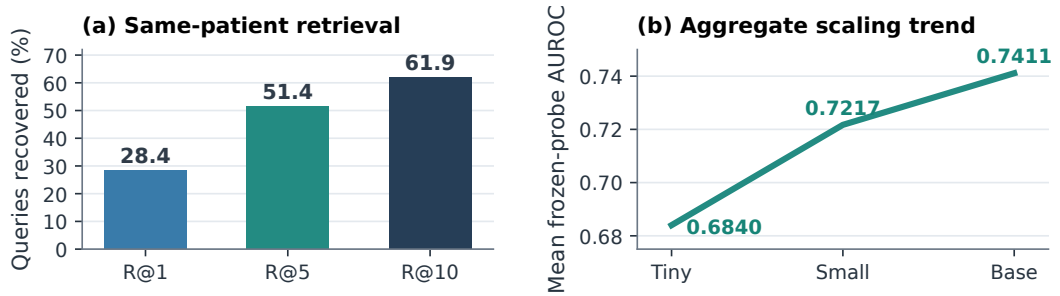

Figure 3: Representation evidence. (a) Same-patient retrieval using PAT. (b) Frozen PAT-probe mean AUROC across three encoder configurations.

Table 3: Frozen-encoder AUROC across seven tasks, grouped by module. Δ is relative to Full.

Configuration	Mean	Δ
Full PAID	**0.8652**	0.0000
Identity structure		
w/o dual space	0.8084	−0.0568
w/o adversarial loss	0.8378	−0.0274
Cross-modal content		
w/o reconstruction	0.6963	−0.1689
w/o semantic prediction	0.7721	−0.0931
Representation regularization		
w/o patient prior	0.8015	−0.0637
w/o variance/covariance	0.8209	−0.0443
w/o temporal consistency	0.8317	−0.0335

Table 4: Inference-time modality removal. Mean excludes readmission. Δ is relative to observed inputs.

Inputs	Mean	Δ
Observed	**0.9084**	0.0000
No CXR	0.9072	−0.0012
No ECG	0.9056	−0.0028
No text	0.8911	−0.0173
Tab.+text	0.9011	−0.0073
Tab. only	0.8752	−0.0332

Removing CXR or ECG has little effect. Tabular plus text retains 0.9011 AUROC, while removing text alone costs 0.0173. These interventions change inputs without retraining.

versus random references of 0.03%, 0.16%, and 0.32% (Figure 3a). The gains of over two orders of magnitude support cross-admission consistency. Appendix A.9 gives the retrieval details and complete results.

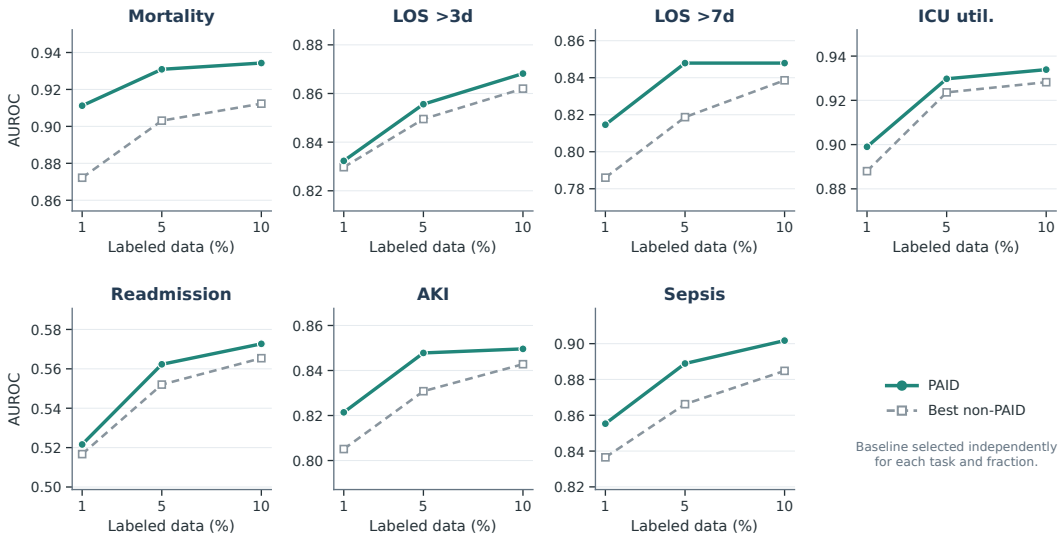

Figure 4: Few-shot AUROC. Each task/fraction uses its strongest reported comparator. Stars mark PAID+XGBoost ensembles. Values are single-run estimates.

Model capacity. Tiny, Small, and Base achieve frozen PAT-probe mean AUROCs of 0.6840, 0.7217, and 0.7411 (Figure 3b). Larger configurations yield stronger frozen representations across these training settings.

Label efficiency (Q3). Figure 4 compares seven tasks at three label fractions. The PAID systems lead all 21 settings, with mean margins of 0.0173, 0.0170, and 0.0106 at 1%, 5%, and 10% labels. The broad gains support transfer to small annotated cohorts. Appendix A.11 gives the complete results and evaluation cohorts. Downstream readouts are specified in Appendix A.3.

Representation specialization (Q4). With common feature-mask training, PAT gives the highest R@5. Combining PAT and ADM improves both clinical summaries, and adding σ gives their highest values. Appendix A.8 provides the settings and endpoint-level results.

Table 5: Representation specialization on MIMIC after clinical fine-tuning. Acute mean excludes readmission.

Representation	R@5	Acute mean	Readm.
PAT	**0.2966**	0.85961	0.6010
ADM	0.1422	0.84612	0.5912
PAT + ADM	0.2942	0.85977	0.6013
PAT + ADM + σ	0.2937	**0.85983**	**0.6016**

5 Conclusion

PAID combines asymmetric identity constraints on PAT and ADM with cross-modal learning to represent recurrent admissions under incomplete observations. Availability-aware attention, cross-modal reconstruction, and structured uncertainty permit learning across visits with different observed modalities. PAID obtains the highest AUROC on all 12 MIMIC-IV and eICU endpoints, improves every evaluated few-shot setting, and retrieves same-patient admissions far above random ranking. Task-dependent PAT/ADM specialization further supports patient linkage as a useful resource when outcome labels are limited. These findings establish recurrent admissions as a practical source of supervision for multimodal clinical representations.

AI USE STATEMENT

Generative AI tools assisted with literature search, code development, and language editing. Reported numerical results are drawn from experimental records.

ETHICS STATEMENT

This study concerns secondary analysis of deidentified clinical records from MIMIC-IV and eICU. Patient linkage is used within each dataset to construct admission pairs and evaluate representations. Patient and admission identifiers remain within the restricted data environment. Cohort selection, documentation, and modality availability may affect generalization. Clinical deployment requires prospective validation.

REPRODUCIBILITY STATEMENT

Appendix A.1 describes inputs and cohorts. Appendices A.3–A.6 describe evaluation settings and training objectives. The appendix also provides complete component ablations, endpoint-level readout comparisons, retrieval metrics, and few-shot results.

REFERENCES

Dara Bahri, Heinrich Jiang, Yi Tay, Donald Metzler, and Da-Cheng Juan. Scarf: Self-supervised contrastive learning using random feature corruption. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pp. 301–318, 2016.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):e1009862, 2022. doi: 10.1371/journal.pcbi.1009862.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam H. Shah, et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7:171, 2024. doi: 10.1038/s41746-024-01166-w.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020a.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020b.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alyssa Gayles, Ammar Shammout, Steven Horng, Tom J. Pollard, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: A retrospective modelling study. *The Lancet Digital Health*, 6(4): e281–e290, 2024. doi: 10.1016/S2589-7500(24)00025-6.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical ModernBERT: An efficient and long context encoder for biomedical text, 2025.

Yikuan Li, Shishir Rao, Jose Robert Solares, Arnuad Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajing Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6: 26094, 2016.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7:119–130, 2025. doi: 10.1038/s42256-024-00965-w.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo Anthony Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc V. Le, Kurt Litsch, Alexander Mossin, Jared Tansuwan, De Wang, James Wexler, Jim Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinath Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeff Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018.

Laila Rasmy, Yu Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533a0.

Sonish Sivarajkumar, Hang Zhang, Yuelyu Ji, Maneesh Bilalpur, Xizhi Wu, Chenyu Li, Min Gu Kwak, Shyam Visweswaran, and Yanshan Wang. A multimodal generative model for structured and unstructured electronic health records. *npj Health Systems*, 3:46, 2026. doi: 10.1038/s44401-026-00095-y.

Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5:149, 2022. doi: 10.1038/s41746-022-00689-4.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective patient representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021. doi: 10.1016/j.jbi.2020.103637.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam H. Shah. MOTOR: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations*, 2024.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHR data. In *International Conference on Learning Representations*, 2025.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

A Appendix

A.1 Inputs and patient cohorts

An admission is represented by five feature blocks in MIMIC. Tokenizers map each block to the encoder width. The number of tokens therefore depends on the modality inventory rather than the dimension of a raw feature block. The same precomputed modality features supply PAID and the baseline implementations. XGBoost concatenates the feature blocks, while the neural methods apply their respective adapters or tokenizers. These comparisons therefore share the upstream extraction pipeline. Training and readout settings are described in Appendix A.3.

Table 6: MIMIC feature inventory. The concatenated input has 1,681 entries. The eICU configuration uses tabular, time-series, and text inputs.

Modality	Dimension	Input representation
Tabular	51	Structured admission features
Time series	64	Summaries of clinical measurements
Chest radiograph	768	RAD-DINO features
Clinical text	768	Clinical ModernBERT features
ECG	30	NeuroKit2 waveform summaries

The implementation represents unavailable feature blocks by zeros and treats the tabular block as present. Availability for the other blocks is derived from their nonzero entries. This convention requires an explicit distinction between a missing block and a genuinely observed all-zero block. Each modality is mapped to a d-dimensional token through a modality-specific Linear–LayerNorm–GELU–Linear network. The dispersion component is computed before adding modality-type embeddings and contributes to the scalar score used for patient-identity matching.

Table 7: Patient counts in the recorded splits. "Recurrent training" is the pool for patient-pair pretraining. The recurrent test group also appears in clinical evaluation.

Patient group	MIMIC	eICU
All patients	180,733	139,367
Recurrent patients	79,535	38,483
Recurrent training	73,569	35,596
Recurrent test	5,966	2,887
Single-admission patients	101,198	100,884
Single-admission downstream training	93,609	93,318
Single-admission downstream test	7,589	7,566

The MIMIC patient map contains 431,231 admissions. All map to the MIMIC-IV hospital admissions metadata used to establish within-patient timestamps. The clinical test cohort combines 25,117 admissions from recurrent test patients with 7,589 single-admission test patients. The retrieval cohort uses the former group. Counts refer to the cohorts used for these experiments, not the size of an entire public database release.

Patient splits are created before pairing admissions. A positive pair uses two different admissions of one training patient. Patients appear only once per minibatch. With more than two admissions, the current sampler chooses an earlier index uniformly from all but the last admission, then chooses a later index uniformly from the remaining later admissions.

A.2 Endpoint labels and evaluation cohorts

The clinical evaluation comprises binary retrospective classification tasks. Table 8 gives the test-cohort size, positive count, and prevalence for each endpoint. Readmission is evaluated on its eligible subset of admissions.

Table 8: Recorded test-set sizes and positive counts. Readmission uses a different eligible subset.

Dataset	Endpoint	Admissions	Positive	Prevalence
MIMIC	Mortality	32,706	670	2.05%
	LOS >3 days	32,706	14,661	44.83%
	LOS >7 days	32,706	5,414	16.55%
	ICU utilization	32,706	4,955	15.15%
	Readmission	12,853	3,486	27.12%
	AKI	32,706	6,319	19.32%
	Sepsis	32,706	1,175	3.59%
eICU	Mortality	15,026	1,293	8.61%
	LOS >3 days	15,026	9,924	66.05%
	LOS >7 days	15,026	4,522	30.09%
	Sepsis	15,026	1,857	12.36%

All endpoints follow a common binary labeling convention: a positive label ($y = 1$) indicates that the clinical record meets the specified outcome criteria, whereas a negative label ($y = 0$) indicates that it does not. Within each dataset, the same labeling rules and cohort eligibility criteria are applied across all compared methods. In Table 8, positive counts correspond to $y = 1$, negative counts equal the total number of evaluated admissions minus the positive counts, and prevalence is the proportion of positive admissions.

A.3 Baselines and the scope of the comparison

XGBoost uses concatenated admission features. MLP and HAIM provide supervised feature-fusion comparisons. AutoEncoder, SimCLR, and SCARF use reconstruction or synthetic views, with SCARF operating on the tabular feature block. Med-BERT-style masked reconstruction and CLMBR-style next-admission prediction are adapted to the pre-extracted admission features. SupCon defines positive pairs using mortality labels.

The clinical tables compare task-specific prediction systems using the common feature-extraction pipeline and patient-disjoint test cohorts. Training pools and readout selection are method-specific. A controlled fine-tuning comparison examines the contributions of the individual readouts (Appendix A.8).

MUSE. MUSE[†] Wu et al. (2024) uses a five-modality adaptation of the patient–modality graph and is trained independently with Adam (learning rate 2×10^{-4}, batch size 512, dropout 0.25, patient-level validation fraction 0.1, seed 42), with random modality-edge deletion disabled and the final checkpoint used for evaluation.

Few-shot readouts. The PAID column in Appendix A.11 reports downstream prediction systems built from PAID features. At 10% labels for LOS >3 days and at 5% and 10% labels for ICU utilization, the readout combines the PAID-feature predictor with a raw-feature XGBoost predictor, with the combination selected on validation data. Other entries use PAID-feature predictors. Thus the few-shot comparison evaluates the resulting prediction systems under limited outcome supervision.

A.4 Role of the scalar score in identity matching

The scalar σ_i adjusts the identity-matching logit to the evidence available in admission i. Equations 5–6 combine dispersion among visible modality tokens, a learned patient-context term, and nonnegative contributions from absent modalities. Dispersion is computed before adding modality-type embeddings, with $\epsilon = 10^{-8}$. The context and missingness heads use two linear layers with GELU and hidden widths $d/4$ and 32, respectively. With one visible modality, the dispersion term equals $\sqrt{\epsilon}$, while context and missingness continue to contribute to the total score.

For a pair with label $y \in \{-1, +1\}$, write its logistic loss as $\ell_y(s) = \log(1 + e^{-ys})$. Differentiating Eq. 7 yields

$$\frac{\partial \ell_y}{\partial \sigma_i} = \gamma e^t y \ \mathrm{sigmoid}(-ys). \tag{11}$$

A larger scalar lowers the matching logit at a fixed cosine similarity. Consequently, positive pairs favor a smaller penalty and negative pairs favor a larger penalty. The class-balanced weights give positive and negative pairs equal aggregate weight in a minibatch. The learned scale and bias are initialized at $e^t = 2$ and $b = -5$, with $\gamma = 0.1$. This formulation learns an evidence-dependent adjustment to patient matching. Appendix A.12 examines its empirical response to modality coverage.

A.5 Masked cross-modal reconstruction

The reconstruction objective trains the encoder to recover an observed modality from the remaining visible evidence. Unavailable and artificially masked modalities are excluded from the encoder's attention keys and values. Before decoding, the encoder output at each artificially masked position is replaced by a learned mask token. Reconstruction at that position therefore relies on visible encoder outputs and the mask token, preventing direct copying of the target through its residual stream.

The decoder projects encoder outputs to width $d/2$, applies two Transformer blocks, and maps each output to the corresponding modality's feature dimension. Equation 9 normalizes the prediction and target within each sample. The loss averages over artificially masked, observed targets within each modality and then over modalities with valid targets. The two admission views contribute equally. Naturally unavailable modalities are excluded from the reconstruction targets.

The masking probabilities for tabular, time-series, CXR, text, and ECG features are $(0, 0.6, 0.6, 0.3, 0.4)$ before the visibility constraint is applied. Tabular features remain visible, together with at least one other observed modality when available. This preserves conditioning information while exposing the encoder to varied combinations of missing modalities.

A.6 Auxiliary objectives and gradient directions

The objectives are organized by training role in the main text. This grouping makes the main architectural contrast clear. It does not imply that removing one loss removes every dependence on patient identity. The semantic predictor, patient prototype, temporal score regularizer, and paired sampler all retain information about the longitudinal relation.

For projected PAT vectors $z_1, \ldots, z_N$ from both views, let $\widetilde{Z}$ be their centered matrix, s_j the sample standard deviation of coordinate j, and $C = \widetilde{Z}^\top \widetilde{Z}/(N-1)$. The implemented variance/covariance regularizer is

$$\mathcal{L}_{\mathrm{vic}} = \frac{1}{D}\sum_j [1 - s_j]_+ + \frac{0.04}{D(D-1)}\sum_{j \neq k} C_{jk}^2. \tag{12}$$

It acts on the unnormalized projection outputs. There is no additional pairwise invariance term in this implementation. Normalization for the identity logit is performed inside the contrastive loss.

The patient prototype is $\pi_p = W E_p + c$, where $E_p \in \mathbb{R}^{64}$ is a training-patient lookup. For one view its loss is

$$\mathcal{L}^{(v)}_{\mathrm{prior}} = \tfrac{1}{2}\left[\mathrm{MSE}(h^{\mathrm{pat}}, \mathrm{sg}(\pi_p)) + \mathrm{MSE}(\pi_p, \mathrm{sg}(h^{\mathrm{pat}}))\right]. \tag{13}$$

The trainer averages this quantity over both views. Stop-gradient sg separates the two update directions. Neither the patient lookup nor a patient identifier is needed for inference on a new patient. The representation is computed from that admission's features.

The semantic head has eight outputs: atelectasis, cardiomegaly, consolidation, edema, no finding, pleural effusion, pneumonia, and pneumothorax. Binary cross-entropy is computed only where a target is available. Uncertain or absent targets are masked in the cache. Early PAT predicts the later admission's target when available, otherwise its own admission's target. CXR is not forcibly hidden in every such example. Accordingly, this objective is not always a prediction solely from non-CXR context and is not free of additional semantic supervision.

The temporal penalty is

$$\mathcal{L}_{\mathrm{temp}} = \mathbb{E}_{(e,l)}[\sigma_l - \sigma_e]_+. \tag{14}$$

It favors a score that does not increase with admission order. There is no clinical guarantee that later admissions are less uncertain: disease and available evidence both change. The legacy integer-only patient map does not establish temporal order. New temporal analyses require admission timestamps joined to the same patient map. Numeric identifier order is not used as a proxy for elapsed time.

For ADM, the matcher parameters minimize the symmetric cross-entropy in Eq. 4. The encoder receives its negative gradient, multiplied by $\alpha_e = 2\min(1, e/10)$ at training epoch e. The outer objective weight is 0.2. The learned temperature is constrained by $\max(|\theta|, 0.01)$. Failure of this particular matcher to identify patients does not establish that all possible adversaries would fail.

Optional Stage II freezes the encoder and trains a newly initialized score head. With missingness count $c_i = M - \sum_m a_{im}$, the intended objective combines the temporal penalty, a squared context-score penalty, and a missingness ranking term, estimated from up to 1,024 sampled pairs per batch:

$$\mathcal{L}_{\mathrm{II}} = 0.5\mathcal{L}_{\mathrm{temp}} + 0.1\mathbb{E}_i[(\sigma_i^{\mathrm{ctx}})^2] + 0.3\mathbb{E}_{i,k}\mathbf{1}(c_i > c_k)[\sigma_k^{\mathrm{miss}} - \sigma_i^{\mathrm{miss}}]_+. \tag{15}$$

This stage changes the score readout without changing frozen PAT or ADM.

A.7 Component sensitivity grouped by training role

Table 3 reports the measured component study in the main text. The detailed grouping below removes one item at a time. We group rows by architectural or training role to distinguish the central design from supporting losses. Rows within a group are separate runs, not joint removal of the entire group.

Table 9: Component sensitivity grouped by role. The deltas are computed relative to this study's Full score.

Role	Removed component	Mean AUROC	Δ
Reference	None	0.8652	0.0000
Identity structure	ADM and matcher	0.8084	−0.0568
	Adversarial loss	0.8378	−0.0274
Content learning	Masked reconstruction	0.6963	−0.1689
	Semantic prediction	0.7721	−0.0931
Regularization	Patient prototype	0.8015	−0.0637
	Variance/covariance	0.8209	−0.0443
	Temporal score penalty	0.8317	−0.0335

Reconstruction and semantic prediction have the largest drops in this comparison. Individual-removal effects also depend on interactions with the retained objectives and need not be additive.

The modality-removal study in the main text is a different intervention: it changes inference-time observations rather than removing a training objective. Its reference mean is 0.9084.

A.8 Controlled fine-tuning of patient and admission readouts

We compare readouts using identical input features, endpoint labels, and patient assignments within each study. The six non-readmission endpoints use 32,706 test admissions, and readmission uses the 12,853 admissions in Table 1. Ten percent of the training patients are reserved for validation, with seed 42. Checkpoints are selected by the mean validation AUROC across seven endpoints.

Training settings. All four readouts use the same pretrained Small encoder and a 20+20 epoch training budget. The initial stage fine-tunes PAT and ADM separately for 20 epochs using AdamW, batch size 512, weight decay 0.01, encoder learning rate 10^{-5}, and head learning rate 10^{-3}. The prediction head has width 256 and dropout 0.1. The first two epochs train only the head, followed by joint fine-tuning with cosine learning-rate decay.

Feature-mask refinement. PAT and the combined readouts start refinement from the selected PAT model, while ADM starts from its selected ADM model after the initial training stage. The combined

heads copy the PAT prediction weights and initialize the additional ADM and σ weights to zero. PAT and ADM are normalized separately before concatenation, while σ is standardized using training-data moments.

Refinement lasts 20 epochs with encoder and head learning rates of 5×10^{-6} and 2×10^{-4}. The common local active-masking strategy is selected by average validation AUROC across the PAT-based readouts and applied unchanged to ADM. Each readout uses its best seven-task validation checkpoint, including the initialization.

An augmented view masks 10% of the nonzero coordinates in one observed non-tabular modality for up to half of the training admissions. Coordinates are ranked by the absolute product of the input and its supervised-loss gradient. The objective combines clean-view and augmented-view binary cross-entropy with weights 0.75 and 0.25, plus prediction consistency with weight 0.1. Masking ramps up after the first two epochs and is removed for the final two epochs. Validation and testing use the observed inputs. Table 10 reports the complete clinical results.

Retrieval after clinical fine-tuning. R@5 in Table 5 uses each readout's fine-tuned encoder on the 25,117-admission retrieval cohort described in Appendix A.9. PAT and ADM use their respective encoder outputs. Combined readouts concatenate these outputs, with the scalar appended for PAT+ADM+σ. Vectors are normalized to unit length and ranked by cosine similarity, excluding the query admission itself. The same selected model supplies each row's retrieval and clinical results.

Table 10: Endpoint-level AUROC after common feature-mask refinement. The six-task mean excludes readmission and reproduces Table 5.

Endpoint	PAT	ADM	PAT+ADM	PAT+ADM+σ
Mortality	0.9262	0.9213	**0.9265**	**0.9265**
LOS >3 days	0.8136	0.7822	0.8139	**0.8140**
LOS >7 days	0.8214	0.8023	0.8217	**0.8218**
ICU utilization	0.8999	0.8907	0.8999	**0.9000**
AKI	**0.8188**	0.8087	**0.8188**	**0.8188**
Sepsis	0.8778	0.8716	**0.8779**	**0.8779**
Readmission	0.6010	0.5912	0.6013	**0.6016**
Six-task mean	0.85961	0.84612	0.85977	**0.85983**
Seven-task mean	0.82267	0.80969	0.82285	**0.82294**

The combined readouts improve both the six-task mean and readmission AUROC over PAT. PAT+ADM+σ attains means of 0.85983 and 0.82294 over six and seven endpoints, respectively. These values describe the shared refinement study in this section.

A.9 Patient retrieval and model capacity

Cross-admission patient retrieval. The retrieval cohort contains 25,117 admissions from 5,966 test patients. Each admission serves as a query, and all other admissions form its candidate set. Candidates are ranked by cosine similarity of PAT vectors. A hit occurs when at least one of the top K candidates belongs to the query patient. Recall@K is averaged over all queries.

Table 11: Same-patient retrieval using PAT. Recall is reported as a proportion, consistent with Figure 3a.

Representation	Recall@1	Recall@5	Recall@10
PAT	0.2841	0.5142	0.6185
Random reference	0.0003	0.0016	0.0032

PAT retrieves another admission of the same patient within the first five candidates for 51.4% of queries. The mean reciprocal rank is 0.3948, the median rank is 5, and the mean rank is 58.58. Together with the large margin over random ranking, these results demonstrate cross-admission identity consistency in the learned patient representation.

Model capacity. The capacity comparison in Figure 3b gives frozen PAT-probe mean AUROCs of 0.6840, 0.7217, and 0.7411 for Tiny, Small, and Base, respectively. The larger configurations yield stronger frozen representations across these training settings.

A.10 eICU ENDPOINT INTERPRETATION

Table 2 compares eight prediction systems across five endpoints. XGBoost is the strongest comparator on all five columns. The four clinical classification endpoints use 15,026 stays, with outcome counts given in Table 8.

The exploratory subsequent-stay endpoint uses 3,718 stays, including 2,274 positives (61.16%). It is analyzed separately from the four clinical classification endpoints and is excluded from the abstract's endpoint count. These results describe retrospective classification and do not establish prospective prediction of a future ICU admission.

Sepsis AUROC is high for several methods, including the autoencoder baseline. This endpoint evaluates retrospective diagnosis classification. Its scores do not quantify performance before clinical recognition or treatment.

A.11 COMPLETE FEW-SHOT RESULTS

Table 12 reports all seven endpoints at 1%, 5%, and 10% labeled training data, together with their absolute AUROC gains. The patient-disjoint test cohorts remain fixed across label fractions: 7,589 single-admission patients for the six non-readmission endpoints and 12,853 eligible admissions for readmission. The corresponding labeled training counts are 936, 4,680, and 9,361 for non-readmission endpoints, and 123, 613, and 1,226 for readmission. Downstream readout settings are specified in Appendix A.3.

Table 12: Complete few-shot AUROC on the MIMIC test cohorts. Best comparator is the strongest non-PAID result for each endpoint and label fraction. Δ is PAID system minus best comparator. The PAID-system readouts are specified in Appendix A.3.

Endpoint	Labels	PAID system	Best comparator	Δ
Mortality	1%	**0.9112**	0.8722	+0.0390
	5%	**0.9309**	0.9031	+0.0278
	10%	**0.9343**	0.9123	+0.0220
LOS >3d	1%	**0.8323**	0.8297	+0.0026
	5%	**0.8556**	0.8495	+0.0061
	10%	**0.8682**	0.8620	+0.0062
LOS >7d	1%	**0.8146**	0.7860	+0.0286
	5%	**0.8479**	0.8187	+0.0292
	10%	**0.8479**	0.8386	+0.0093
ICU utilization	1%	**0.8990**	0.8880	+0.0110
	5%	**0.9297**	0.9236	+0.0061
	10%	**0.9339**	0.9282	+0.0057
Readmission	1%	**0.5216**	0.5167	+0.0049
	5%	**0.5623**	0.5520	+0.0103
	10%	**0.5727**	0.5654	+0.0073
AKI	1%	**0.8214**	0.8051	+0.0163
	5%	**0.8478**	0.8308	+0.0170
	10%	**0.8496**	0.8428	+0.0068
Sepsis	1%	**0.8553**	0.8365	+0.0188
	5%	**0.8889**	0.8662	+0.0227
	10%	**0.9017**	0.8848	+0.0169
Mean	1%	**0.8079**	0.7906	+0.0173
	5%	**0.8376**	0.8206	+0.0170
	10%	**0.8440**	0.8334	+0.0106

The PAID systems achieve higher AUROC in all 21 settings, with mean gains of 0.0173, 0.0170, and 0.0106 at 1%, 5%, and 10% labels, respectively. At 1% labels, the largest gains occur for mortality (0.0390), LOS $>$7 days (0.0286), and sepsis (0.0188). Positive gains across all seven endpoints support the transferability of PAID representations when outcome annotations are scarce.

A.12 MATCHING SCORE AND MODALITY COVERAGE

The score analysis contains 5,966 sampled admission pairs. The score has Spearman correlation -0.5086 with the number of observed modalities and 0.5086 with the number of missing modalities. Since the score explicitly includes absent-modality terms and can be trained to rank missingness, these correlations mainly describe its structural response.

For each pair, let $\Delta\sigma = \sigma_{\text{second}} - \sigma_{\text{first}}$. Table 13 conditions this difference on relative modality coverage. Pair indices denote sampling order.

Table 13: Score change by relative modality coverage in sampled pairs. Pair shares are given in parentheses.

Second admission's coverage	Pairs	$P(\Delta\sigma \leq 0)$	Mean $\Delta\sigma$
More	1,739 (29.15%)	66.53%	-0.0831
Same	2,441 (40.92%)	49.16%	-0.0010
Fewer	1,786 (29.94%)	32.64%	$+0.0852$