MLPA: MULTI-MODAL REPRESENTATION LEARNING FROM INCOMPLETE HEALTH RECORDS VIA MASKED LATENT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Integrating multi-modal Electronic Health Records (EHR) is difficult. Clinical streams vary widely in length, missingness is often clinically meaningful rather than random, and labeled outcomes are scarce. High-volume streams such as medication logs can dominate cross-modal attention and crowd out sparser signals, a failure mode we term modality washout. The objectives typically used for multi-modal integration add to the problem. Generative losses spend capacity on reconstructing high-entropy raw inputs, while supervised losses shape representations around a single downstream target. We introduce the Multi-modal Latent Predictive Architecture (MLPA), a self-supervised framework that learns patient representations from incomplete clinical records by predicting masked targets in uni-modal latent space. MLPA routes cross-modal exchange through a shared bottleneck transformer and represents missing modalities with learned empty-state tokens. We evaluate MLPA on 81,664 MIMIC-IV admissions and 188,094 eICU stays. On MIMIC-IV, MLPA matches a multi-modal masked autoencoder with 42.4% fewer parameters. On eICU, it achieves higher mortality AUPRC than the autoencoder (0.430 vs. 0.403). Without labels at pre-training, MLPA outperforms the semi-supervised method MUSE+ on in-hospital mortality in both datasets (eICU AUPRC 0.430 vs. 0.388). In eICU, this advantage holds for stays with one missing modality. MLPA also forms more compact clusters than MUSE+ for outcomes that neither model was trained on. An attention analysis shows that unconstrained cross-attention gives 47% of its attention to medication tokens, while the bottleneck keeps attention balanced across streams.

1 INTRODUCTION

Clinical decision-making depends on heterogeneous patient data. In routine care, clinicians combine observations from several modalities, such as past diagnoses, administered medications and imaging results, to assess disease progression and guide treatment (Acosta et al., 2022; Kline et al., 2022). Multi-modal representation learning aims to reproduce this synthesis. It extracts patient representations that support downstream clinical tasks, such as in-hospital mortality prediction and disease classification.

The challenges of multi-modal learning from Electronic Health Records (EHR) are not unique to this domain. Modalities of very different length also occur in audio-visual data (Nagrani et al., 2021). Missing modalities also occur in vision-language data (Ma et al., 2021; 2022). Scarce labels motivate self-supervised learning in many domains (He et al., 2022; Baevski et al., 2022). EHR data, however, combine all three challenges in a single setting and make each of them more severe.

First, clinical streams differ strongly in volume. Medication logs contain about an order of magnitude more tokens per admission than procedure codes or radiology reports, and demographics contain only two (Section 4.6). Prior work on modality imbalance studies how strongly or how quickly modalities drive learning (Wang et al., 2020). Other work studies models that over-rely on a subset of modalities (Javaloy et al., 2022; Wu et al., 2024). We address a volume-driven variant

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.

of this problem. When attention is unconstrained, long streams can take most of the cross-modal capacity and suppress short but informative streams. We term this failure mode *modality washout*.

Second, missingness in EHR is informative. In many multi-modal settings, a modality is missing because of a technical failure. In EHR, a radiology report is often absent because a physician decided that imaging was not needed (Lipton et al., 2016; Che et al., 2018). A missing observation is therefore information in itself. Imputation removes this signal (Groenwold, 2020), and explicitly modeling the missingness mechanism (Ipsen et al., 2021) is hard across heterogeneous streams. Third, labeled clinical outcomes are scarce, because labels require clinical expertise or follow-up. This limits supervised integration on large, mostly unlabeled cohorts.

Existing methods address parts of this problem (Section 2). Generative methods reconstruct raw inputs. This requires a decoder and a vocabulary-sized output head for each modality, and spends capacity on high-entropy details such as billing artifacts and boilerplate text. Supervised and semi-supervised methods, such as MUSE+ (Wu et al., 2024), learn from downstream labels. Their representations are therefore shaped around the tasks they are trained on. Architectures with unconstrained cross-modal attention let the longest streams dominate.

We introduce the Multi-modal Latent Predictive Architecture (MLPA; Figure 1). Each modality is encoded by a frozen uni-modal encoder. A Multi-modal Bottleneck Transformer (MBT) (Nagrani et al., 2021) processes each stream separately and limits cross-modal exchange to a small set of shared tokens. These tokens are pooled into a single fused representation Z_f. During self-supervised pre-training, MLPA predicts the frozen encoder outputs at masked positions of each observed modality, using only Z_f. All information needed for this prediction must therefore pass through the fused representation. Missing modalities are represented by learned empty-state tokens instead of imputed values. Our contributions are:

- *Self-supervised masked latent prediction for incomplete EHR.* MLPA combines fixed latent targets, prediction from a single fused representation, bottleneck routing and learned empty-state tokens. Because the targets come from frozen encoders, MLPA does not need negative pairs, a teacher network or collapse regularization.
- *A controlled comparison of prediction targets.* We compare MLPA with a masked autoencoder that shares its backbone, masking and missingness handling. MLPA matches it on MIMIC-IV and achieves higher mortality AUPRC on eICU, with 42.4% fewer trainable parameters.
- *Self-supervised versus semi-supervised representations.* Without labels at pre-training, MLPA outperforms MUSE+ on in-hospital mortality in both datasets, including eICU stays with one missing modality, and matches it on phenotyping. It also forms more compact clusters for outcomes that no model was trained on. An attention analysis shows that the bottleneck keeps cross-modal attention balanced, while unconstrained cross-attention concentrates on medication tokens.

2 Related Work

Incomplete multi-modal EHR. Imputation-based methods reconstruct missing modalities from observed ones, for example with cross-modal autoencoders (Ngiam et al., 2011) or Bayesian meta-learning (Ma et al., 2021). Direct prediction methods work with incomplete inputs. They use transformers (Ma et al., 2022), disentangled fusion (Yao et al., 2024), generative models of heterogeneous EHR (He et al., 2024), or graphs. GRAPE (You et al., 2020) links patients and modalities in a bipartite graph. M3Care (Zhang et al., 2022) imputes task-relevant information from similar patients. MUSE (Wu et al., 2024) adds mutual-consistent contrastive learning on the patient-modality graph, and MUSE+ extends its training to unlabeled patients. Recent work adds uncertainty-aware message passing (Gong et al., 2026) or causal debiasing (Zhu et al., 2025) to these graph models.

MLPA differs from these graph models in three ways. First, it uses no labels at pre-training, so one representation serves all tasks. Second, a patient's representation depends only on that patient's record, while MUSE builds its patient-modality graph over each batch, including at inference. Third, cross-modal exchange passes through a fixed set of bottleneck tokens instead of message passing. The closest work is CRL-MMNAR (Liang et al., 2025). It treats missingness patterns as signal, fuses gated modality embeddings with attention, and reconstructs held-out modality embeddings from the

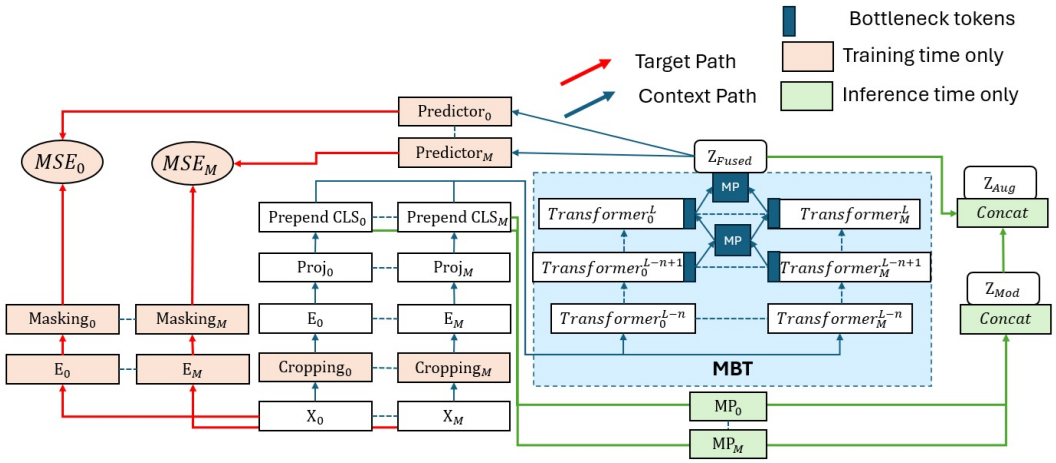

Figure 1: **Overview of MLPA. (a)** Self-supervised training. The context path (blue) encodes the unmasked positions of each observed modality with a frozen encoder E_m and a projection P_m. The empty-state token CLS_m is prepended, and the MBT combines all streams into Z_f. The predictors Pred_m map Z_f and positional mask tokens to the frozen encoder outputs at the masked positions (target path, red). **(b)** MBT. Streams are processed independently for $L - n$ layers. In the last n layers, they exchange information only through the bottleneck tokens $B^{(l)}$, which average the per-stream candidates $\hat{B}_m^{(l)}$. **(c)** Inference. Uncropped inputs pass through the same encoders, projections and MBT. Z_f is the patient representation.

remaining modalities. It is trained jointly with supervised multi-task heads. In contrast, MLPA predicts masked positions within each modality from a single bottleneck representation, represents missingness with empty-state tokens, and uses no labels at pre-training.

Latent prediction. data2vec (Baevski et al., 2022) and I-JEPA (Assran et al., 2023; LeCun, 2022) predict latent targets produced by a teacher network that is updated during training. This requires mechanisms that prevent representation collapse (Bardes et al., 2021). Other methods predict fixed targets, such as hand-crafted features (Wei et al., 2022) or the features of a frozen teacher (Peng et al., 2023). These methods avoid collapse by construction. MLPA belongs to this second group, since its targets are the outputs of frozen uni-modal encoders. MLPA also differs from JEPA-style models in two ways. Its predictor receives a single pooled multi-modal representation instead of context tokens. It also operates on records in which entire modalities can be missing. Joint masked objectives over several modalities (Singh et al., 2022) use both intra- and cross-modal redundancy. MLPA applies the same idea with latent targets.

Cross-modal routing. Early fusion applies self-attention over the concatenation of all streams, at a cost that is quadratic in the total number of tokens. MulT (Tsai et al., 2019) uses pairwise cross-modal attention, which grows with the number of modality pairs. Perceiver-style models (Jaegle et al., 2022) use a fixed set of latent queries that attend over the flattened input at linear cost. Their attention can, however, concentrate on the longest streams (Section 4.6). The MBT (Nagrani et al., 2021) processes streams separately and restricts cross-modal exchange to shared bottleneck tokens. MLPA adopts this routing.

3 Multi-modal Latent Predictive Architecture

3.1 Modality-specific representations

Consider a set of M clinical modalities, indexed by $m \in \{1, \ldots, M\}$. Each modality observation X_m is mapped to hidden states by a frozen, modality-specific encoder E_m. Non-sequential modalities are formatted as fixed-length sequences. For example, demographics are mapped to a two-token sequence for age and sex. A learned linear projection P_m maps the encoder outputs to a shared

dimension d:

$$U_m = P_m\big(E_m(X_m)\big) \in \mathbb{R}^{T_m \times d}, \tag{1}$$

where T_m is the sequence length of modality m. To represent missing observations, a learnable empty-state token $\mathrm{CLS}_m \in \mathbb{R}^{1\times d}$ is prepended to each projected sequence:

$$Z_m = [\mathrm{CLS}_m \parallel U_m], \tag{2}$$

where $\parallel$ denotes concatenation along the sequence dimension. When modality m is not observed, $U_m = \emptyset$ and the sequence reduces to $Z_m = [\mathrm{CLS}_m]$. This avoids imputation and per-sample branching. CLS_m is learned end-to-end through the shared objective.

3.2 Multi-modal Bottleneck Transformer and fused representation

MLPA integrates the streams $\{Z_m\}_{m=1}^M$ with an MBT (Nagrani et al., 2021) of L layers. The MBT consists of M parallel transformer encoders. In the first $L-n$ layers, each stream is updated independently. In the final n layers, cross-modal exchange is restricted to K shared bottleneck tokens $B^{(l)} \in \mathbb{R}^{K\times d}$. The initial tokens $B^{(L-n)}$ are learned parameters. At each fusion layer l, the bottleneck tokens are prepended to each stream. Each stream then returns an updated bottleneck candidate $\hat{B}_m^{(l)}$, which is its first K output tokens, and its updated stream tokens:

$$[\hat{B}_m^{(l)} \parallel Z_m^{(l)}] = \mathrm{Transformer}_m^{(l)}\big([B^{(l-1)} \parallel Z_m^{(l-1)}]\big). \tag{3}$$

The bottleneck tokens for the next layer are the average of the candidates over all streams:

$$B^{(l)} = \frac{1}{M}\sum_{m=1}^{M} \hat{B}_m^{(l)}. \tag{4}$$

This average is the only path between modalities. Streams of missing modalities contain only CLS_m, but they still contribute a candidate. This is how missingness reaches the fused representation, and how CLS_m receives gradients. At layer L, the terminal bottleneck tokens $b_k^{(L)}$ are mean-pooled into the fused representation:

$$Z_f = \frac{1}{K}\sum_{k=1}^{K} b_k^{(L)} \in \mathbb{R}^d. \tag{5}$$

3.3 Masked latent prediction

Masking. For each observed modality, the positions $\{1,\dots,T_m\}$ are split into context positions $\mathcal{C}_m$ and target positions $\mathcal{T}_m$. A stochastic operator selects contiguous target segments. The number and length of the segments scale with T_m.

Target path. The full input X_m passes through the frozen encoder, $Z_m^t = E_m(X_m)$. The targets $z_{m,i}^t$ are the encoder outputs at the target positions $i \in \mathcal{T}_m$. CLS_m is not a target.

Context path. Only the context positions are encoded. The result is projected and prepended with the empty-state token:

$$Z_m^{\mathrm{ctx}} = [\mathrm{CLS}_m \parallel P_m\big(E_m(X_{\mathcal{C}_m})\big)]. \tag{6}$$

The MBT maps the context streams of all modalities to Z_f. A masked position can therefore be predicted from the remaining context of its own modality, or from correlated signal in other modalities.

Predictors and objective. Each observed modality $m \in \mathcal{O}$ has a lightweight predictor Pred_m. The predictor receives Z_f as a single-token prefix, followed by one learnable positional mask token per target position:

$$H_{\mathrm{pred},m} = [Z_f \parallel E_{\mathrm{mask}}^{(m)}], \qquad E_{\mathrm{mask}}^{(m)} \in \mathbb{R}^{|\mathcal{T}_m|\times d}. \tag{7}$$

The predictor outputs $\hat{z}_{m,i}$ in the native space of E_m. It therefore maps from the fused multi-modal space to each uni-modal target space. The model is trained by minimizing the mean squared error over observed modalities only:

$$\mathcal{L} = \sum_{m\in\mathcal{O}} \frac{1}{|\mathcal{T}_m|} \sum_{i\in\mathcal{T}_m} \big\|\hat{z}_{m,i} - z_{m,i}^t\big\|_2^2. \tag{8}$$

Missing modalities contribute no loss.

This design has two consequences. First, Pred_m only receives Z_f and position tokens. Any context information that is useful for the targets must therefore pass through Z_f. Because the predictor has low capacity, the MBT must make this information readable. Second, the encoders are frozen, so the targets are fixed and non-degenerate. The loss cannot be reduced by collapsing the representation. MLPA therefore needs no negative pairs, teacher network or variance regularization.

3.4 Inference

At inference, uncropped inputs are encoded, projected and passed through the MBT. The fused representation Z_f is used for all downstream evaluations. The predictors and mask tokens are not used.

4 Experiments and Results

4.1 Experimental Setup

Cohorts and Modalities. We evaluate on 81,664 adult hospital admissions from MIMIC-IV (Johnson et al., 2023) and 188,094 ICU stays from eICU (Pollard et al., 2018). Each record contains up to four modalities: clinical codes, administered medications, clinical text, and demographics (age and sex). In MIMIC-IV, the codes are ICD-10 procedure codes and the text consists of radiology reports. eICU has no procedure codes, so its code stream contains ICD diagnosis codes. Its text comes from the eICU notes table. Radiology reports are missing in 43.9% of MIMIC-IV admissions, and 34.6% of eICU stays miss at least one modality. Input sequences cover each stay from admission to discharge, or to death for patients who die in hospital. Mortality is therefore classified from the complete record, not predicted in advance. We use this task to compare representations under identical inputs. In MIMIC-IV, discharge summaries, post-mortem notes, and retrospective billing flags are removed. Training and test sets are split by patient. Cohort selection, leakage filtering, and missingness patterns are detailed in Appendix G.

Downstream Tasks. We assess representations on two tasks: (i) binary *in-hospital mortality* on both cohorts (prevalence: 5.32% in MIMIC-IV; 9.16% in eICU), and (ii) multi-label *phenotyping* over ten diagnostic groups on MIMIC-IV (macro-averaged AUPRC and AUROC; Appendix B). In MIMIC-IV, diagnosis codes are only used to create the phenotype labels. Phenotyping is omitted on eICU, because diagnosis codes are an input modality there.

Baselines and Representations. The code, medication, and text streams use BERT encoders pretrained with masked language modeling on the training set of each dataset. Demographics use a deterministic encoder. All encoders are frozen during multi-modal training (Appendix G). We compare MLPA (Z_f) with:

- **Masked Autoencoder (MAE):** Shares MLPA's encoders, masking, empty-state tokens, and bottleneck architecture, but reconstructs the masked inputs in raw token space (He et al., 2022). The MAE therefore isolates the effect of predicting in latent space.
- **MUSE+:** A graph-based, semi-supervised multi-modal method (Wu et al., 2024). In its original evaluation, MUSE+ outperformed M3Care, GRAPE, HGMF, MT, SMIL, and CM-AE (Zhang et al., 2022; You et al., 2020; Chen & Zhang, 2020; Ma et al., 2022; 2021; Ngiam et al., 2011). It also remained the strongest baseline for mortality in an independent re-evaluation on different cohorts (Liang et al., 2025). We train a separate MUSE+ model for each task, on the same frozen uni-modal embeddings as MLPA. Fine-tuning the encoders, as in the original work, gives similar results on MIMIC-IV, except for a higher phenotyping AUROC (Appendix C.5). Published MUSE+ results are not comparable to ours, because they use different cohorts, task definitions, and modalities.
- **Uni-Modal and Concatenation:** The frozen encoder outputs, alone and concatenated (late integration).

Evaluation Protocol. MLPA and the MAE are pre-trained once on the training set and then frozen. We evaluate their representations with a two-layer MLP probe, and with a linear probe where noted,

Table 1: **Downstream task performance across cohorts and missingness patterns.** MAE and MLPA are evaluated with an MLP probe on the frozen Z_f. MUSE+ is trained separately for each endpoint and uses its own classifier. Values are mean$_{\pm \text{s.d.}}$ over 10 folds of the held-out test set. *Computed on the stays of each subgroup from the out-of-fold predictions of the all-stays probe. In both datasets, mortality is classified from the complete record (Section 4.1). Trainable parameters on MIMIC-IV: MAE 3.3M, MUSE+ 0.8M, MLPA 1.9M. Bold marks the highest mean in each group.

			Phenotyping		**In-Hospital Mortality**	
Dataset	**Subgroup**	**Model**	AUPRC	AUROC	AUPRC	AUROC
MIMIC-IV	All ($N = 16{,}333$)	MAE	$0.828_{\pm.003}$	$0.821_{\pm.002}$	$0.880_{\pm.037}$	$0.989_{\pm.004}$
		MUSE+	$0.823_{\pm.003}$	$0.817_{\pm.003}$	$0.672_{\pm.042}$	$0.952_{\pm.012}$
		MLPA	$0.825_{\pm.004}$	$0.815_{\pm.003}$	$0.886_{\pm.022}$	$0.987_{\pm.004}$
	All present ($N = 9{,}160$)	MAE	$0.828_{\pm.006}$	$0.810_{\pm.005}$	$0.851_{\pm.036}$	$0.980_{\pm.008}$
		MUSE+	$0.831_{\pm.005}$	$0.815_{\pm.005}$	$0.638_{\pm.095}$	$0.938_{\pm.018}$
		MLPA	$0.826_{\pm.004}$	$0.805_{\pm.004}$	$0.864_{\pm.038}$	$0.977_{\pm.013}$
	Rad. missing ($N = 7{,}173$)	MAE	$0.819_{\pm.005}$	$0.822_{\pm.003}$	$0.867_{\pm.056}$	$0.990_{\pm.004}$
		MUSE+	$0.816_{\pm.005}$	$0.819_{\pm.004}$	$0.706_{\pm.065}$	$0.962_{\pm.014}$
		MLPA	$0.816_{\pm.004}$	$0.816_{\pm.003}$	$0.871_{\pm.053}$	$0.990_{\pm.004}$
eICU	All stays ($N = 37{,}619$)	MAE	–	–	$0.403_{\pm.024}$	$0.844_{\pm.008}$
		MUSE+	–	–	$0.388_{\pm.032}$	$0.818_{\pm.010}$
		MLPA	–	–	$0.430_{\pm.028}$	$0.855_{\pm.008}$
	All present* ($N = 24{,}734$)	MAE	–	–	$0.413_{\pm.044}$	$0.843_{\pm.014}$
		MUSE+	–	–	$0.408_{\pm.043}$	$0.820_{\pm.015}$
		MLPA	–	–	$0.461_{\pm.035}$	$0.860_{\pm.014}$
	1 missing* ($N = 10{,}419$)	MAE	–	–	$0.377_{\pm.044}$	$0.849_{\pm.014}$
		MUSE+	–	–	$0.352_{\pm.032}$	$0.812_{\pm.023}$
		MLPA	–	–	$0.386_{\pm.042}$	$0.854_{\pm.009}$
	2 missing* ($N = 2{,}466$)	MAE	–	–	$0.231_{\pm.082}$	$0.786_{\pm.059}$
		MUSE+	–	–	$0.374_{\pm.120}$	$0.788_{\pm.060}$
		MLPA	–	–	$0.250_{\pm.096}$	$0.802_{\pm.051}$

using 10-fold cross-validation on the held-out test set. We report the mean and standard deviation over folds. MUSE+ is trained on the training set with its own classifier, and evaluated on the same test folds.

4.2 Latent Prediction vs. Raw Input Reconstruction

On MIMIC-IV, MLPA matches the MAE on both tasks (Table 1). For mortality, AUPRC is 0.886 for MLPA and 0.880 for the MAE, and AUROC is 0.987 and 0.989. For phenotyping, AUPRC is similar (0.825 vs. 0.828), and AUROC is slightly lower for MLPA (0.815 vs. 0.821). On the larger, more heterogeneous eICU cohort, MLPA outperforms the MAE on mortality (AUPRC 0.430 vs. 0.403; AUROC 0.855 vs. 0.844). MLPA does not need decoders or vocabulary-sized output heads. This reduces the number of trainable parameters by 42.4%, from 3.3M to 1.9M. The two objectives also produce differently organized embeddings (Section 4.7).

4.3 MLPA Outperforms Endpoint-Trained MUSE+ on Mortality

MUSE+ is trained for each endpoint, while MLPA uses no labels before probing. On in-hospital mortality, MLPA outperforms MUSE+ on both cohorts (Table 1). On MIMIC-IV, AUPRC is 21.4 points higher (0.886 vs. 0.672) and AUROC is 3.5 points higher (0.987 vs. 0.952). On eICU, AUPRC is 4.2 points higher (0.430 vs. 0.388) and AUROC is 3.7 points higher (0.855 vs. 0.818). On multi-label phenotyping, where label diversity is higher, MLPA matches MUSE+ (AUPRC 0.825 vs. 0.823). These results suggest that MLPA learns task-agnostic representations that are also label-efficient.

One possible explanation is the class imbalance of the target, with a mortality prevalence of 5.32% in MIMIC-IV. A model trained for such a target can rely on a few strongly correlated features in the dominant modalities. This may make its representation less robust. MLPA is instead trained with independent per-modality masking. An event whose segment is masked in one stream can remain visible in another, so predicting the frozen uni-modal latents encourages information transfer across modalities. This may help the bottleneck retain information about the patient's state from all streams. The resulting representation transfers well to downstream endpoints, including under a linear probe (Section 4.5).

4.4 Performance Under Modality Missingness

Informative Missingness in MIMIC-IV. In MIMIC-IV, only radiology reports are ever missing (43.9% of admissions). These admissions form a distinct subpopulation, for which no imaging was ordered or reported. They have a higher mortality prevalence (5.91% vs. 4.80%) and a larger share of admissions without a phenotype label (7.96% vs. 1.02%). We therefore compare models within each subset, and not scores across subsets. The ranking of the models on mortality AUPRC is the same in both subsets (Table 1). MLPA matches the MAE, and it outperforms MUSE+ whether radiology reports are present (AUPRC 0.864 vs. 0.638) or missing (0.871 vs. 0.706).

Missing Clinical Streams in eICU. In eICU, the missing modalities include the two most informative streams, diagnosis codes and medications (Table 2). On stays with all modalities present, MLPA outperforms both baselines (AUPRC 0.461 vs. 0.413 for the MAE and 0.408 for MUSE+). With one missing modality, MLPA matches the MAE and outperforms MUSE+ (AUPRC 0.386, 0.377, and 0.352). Appendix E reports results for each missingness pattern.

On stays with two missing modalities (6.6% of the test set), MUSE+ achieves higher AUPRC than MLPA (0.374 vs. 0.250), while AUROC is similar (0.788 vs. 0.802). The MAE, which shares MLPA's masking and empty-state tokens, has a similarly low AUPRC (0.231). This points to the shared handling of absent streams, not to the prediction target. MUSE+ drops graph edges during training, which exposes it to entirely missing streams. MLPA's independent per-modality masking hides segments within streams but never a whole stream, so MLPA only encounters missing streams in naturally incomplete records. Two missing streams are rare in these records. The resulting combination of empty-state tokens may therefore shift the inputs of the bottleneck away from its training distribution. Whole-modality dropout during pre-training could address this.

4.5 Information Retention and Fusion Gains

To test whether Z_f captures cross-modal information, rather than only summarizing the dominant streams, we compare it with the frozen uni-modal embeddings and their concatenation (late integration). All representations are evaluated on records with all modalities present. On MIMIC-IV, we pre-train a separate MLPA model only on admissions with all modalities present (the complete-case model), to match the data available to concatenation. On eICU, we use the main model, pre-trained on all stays.

On MIMIC-IV, medications are the most informative single stream for both tasks, and demographics the least. On eICU, diagnosis codes are the most informative stream for mortality (AUPRC 0.312), and clinical notes the least (0.235). On both datasets, concatenation achieves higher AUPRC than every single stream, which indicates that the modalities carry complementary information.

Z_f achieves higher AUPRC than concatenation for both tasks and both probes, although it is a single 128-dimensional vector. On MIMIC-IV, the difference is largest with the linear probe (mortality 0.812 vs. 0.768). The MLP probe lets concatenation model interactions between features, but Z_f still achieves higher AUPRC (mortality 0.872 vs. 0.849 on MIMIC-IV, and 0.461 vs. 0.400 on eICU). Both representations are computed from the same frozen encoders and evaluated with the same probes. The gain is therefore unlikely to come from the uni-modal encoders or from the capacity of the classifier. The bottleneck processes the full token sequence of each stream and exchanges information between streams. This suggests that it retains dependencies that pooling each stream separately discards.

On MIMIC-IV, the main model, pre-trained on all admissions, matches the complete-case model on the same test patients (phenotyping AUPRC 0.826 for both; mortality 0.864 vs. 0.872; Tables 1

Table 2: **Uni-modal embeddings, late integration, and bottleneck fusion (Z_f) on records with all modalities present.** AUPRC mean$_{\pm \text{s.d.}}$ over 10 folds (MIMIC-IV $N = 9{,}160$; eICU $N = 24{,}734$). Z_f is a single 128-dimensional vector. On MIMIC-IV, MLPA is pre-trained on admissions with all modalities present. On eICU, it is the main model, pre-trained on all stays. Bold marks the highest mean in each column.

	MIMIC-IV				eICU
	Linear probe		MLP probe		MLP probe
Representation	Pheno.	Mortality	Pheno.	Mortality	Mortality
Billing ICD Codes	$0.710_{\pm.006}$	$0.355_{\pm.052}$	$0.742_{\pm.006}$	$0.433_{\pm.059}$	$0.312_{\pm.021}$
Medications	$0.778_{\pm.007}$	$0.724_{\pm.078}$	$0.801_{\pm.006}$	$0.832_{\pm.066}$	$0.306_{\pm.027}$
Clinical Notes / Text	$0.713_{\pm.008}$	$0.397_{\pm.070}$	$0.723_{\pm.007}$	$0.507_{\pm.100}$	$0.235_{\pm.020}$
Demographics	$0.616_{\pm.010}$	$0.087_{\pm.022}$	$0.665_{\pm.009}$	$0.097_{\pm.024}$	$0.275_{\pm.018}$
Concatenation (Late)	$0.807_{\pm.005}$	$0.768_{\pm.064}$	$0.822_{\pm.005}$	$0.849_{\pm.063}$	$0.400_{\pm.037}$
MLPA (Z_f)	$0.817_{\pm.004}$	$0.812_{\pm.068}$	$0.826_{\pm.005}$	$0.872_{\pm.050}$	$0.461_{\pm.028}$

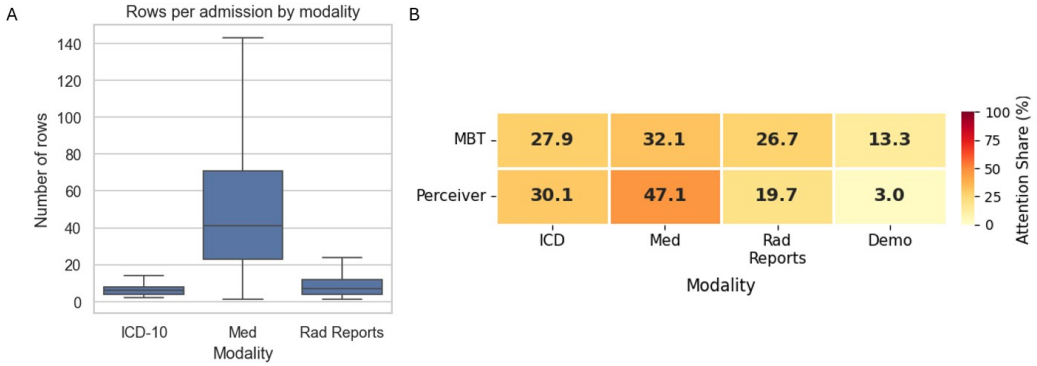

Figure 2: **Token volume and attention allocation on MIMIC-IV. (A)** Token count per admission for the variable-length modalities (demographics always have two tokens). **(B)** MBT: mean attention share of each bottleneck token per modality, across all fusion layers. **(C)** Perceiver: share of total cross-attention per modality, across its two cross-attention layers.

and 2). This suggests that pre-training on incomplete records does not reduce performance on complete ones.

4.6 Attention Allocation: Bottleneck vs. Unconstrained Attention

EHR modalities differ strongly in token volume. Medication logs produce orders of magnitude more tokens than procedure codes or demographics (Figure 2A). In unconstrained cross-attention, this imbalance risks modality starvation, where dense streams dominate the attention.

To test this, we replace the MBT with Perceiver-style cross-attention (Jaegle et al., 2022), in which latent queries attend over the flattened tokens of all streams. Both variants are pre-trained on MIMIC-IV admissions with all modalities present. With a linear probe, the MBT matches the Perceiver on the phenotyping task (AUPRC 0.817 vs.0.815), while shown superior performance on the mortality task (AUPRC 0.812 vs.0.796).

The two variants allocate attention differently (Figure 2B and C). The Perceiver gives 47.1% of its cross-attention to medication tokens and 3.0% to demographics, broadly in line with the token volumes in panel A. Each MBT bottleneck token gives 26–32% of its attention to each clinical stream and 12–15% to demographics. Part of this balance is structural, because the MBT averages the modality-specific bottleneck candidates with equal weight. Attention share is also not the same as contribution to the representation. With four modalities, the balanced allocation does not lead to

Table 3: **Cluster compactness (Davies–Bouldin index) for clinical labels on the MIMIC-IV test set.** Lower is more compact and better separated. Phenotype: ten diagnostic groups. Cardiovascular: four cardiovascular subgroups. †Label that the MUSE+ model was trained on; all other model–label pairs are unsupervised. Single run. Bold marks the lowest value in each column.

Model	Mortality	Phenotype	Readmission	Cardiovascular
MLPA (Z_f)	3.11	**7.23**	**9.30**	**7.39**
MAE (Z_f)	3.86	7.30	10.27	7.81
MUSE+ (mortality-trained)	**3.01**†	14.56	25.34	15.06
MUSE+ (phenotype-trained)	8.57	11.60†	27.23	14.94

a measurable difference in downstream performance. It may, however, help protect against modality starvation when scaling to higher-volume modalities, such as high-frequency physiological signals or long clinical notes.

4.7 Intrinsic Representation Geometry on Unseen Endpoints

To test whether the representation organizes patients beyond the probed tasks, we compute the Davies–Bouldin index (DBI) (Davies & Bouldin, 1979) of the test-set embeddings for four clinical labels: mortality, phenotype group, hospital readmission, and cardiovascular subgroup (Table 3). A lower DBI means more compact and better separated groups. Appendix F describes how labels are assigned and how embeddings are normalized.

MLPA forms more compact clusters than the MAE for all four labels, which suggests that predicting in latent space yields a more structured representation than reconstructing raw inputs. On the labels that a MUSE+ model was trained on, MLPA is similarly compact for mortality (3.11 vs. 3.01) and more compact for phenotypes (7.23 vs. 11.60). On the labels that no model was trained on, both MUSE+ models are much less compact than MLPA. For readmission, DBI is 25.34 and 27.23 vs. 9.30, and for cardiovascular subgroups 15.06 and 14.94 vs. 7.39. DBI measures compactness, not predictive performance. Still, these results suggest that endpoint-specific training organizes the representation around its training label, at some cost to its general clinical utility.

5 Conclusion and Limitations

We introduced MLPA, a self-supervised framework that learns patient representations from incomplete multi-modal EHR records by predicting frozen uni-modal latents through a bottleneck transformer. On two datasets, MLPA matched or outperformed raw-input reconstruction with 42.4% fewer trainable parameters. It outperformed the endpoint-trained MUSE+ on mortality, matched it on phenotyping, and formed more compact clusters for outcomes that no model was trained on. Its fused representation also achieved higher AUPRC than late integration of the same frozen embeddings.

MLPA remained robust when one modality was missing, but trailed MUSE+ in AUPRC when two of three clinical streams were missing in eICU. Whole-modality dropout during pre-training, and a comparison of our empty-state tokens with simpler alternatives such as zero-filling, are immediate next steps. Because input sequences span the whole stay and end at death for non-survivors, our mortality task is retrospective classification, not prospective early warning. Our evaluation is restricted to four modalities, and our attention analysis is descriptive and does not causally isolate cross-modal transfer. Each model was also pre-trained once, so the reported standard deviations capture probe variance only. Future work will validate MLPA on prospective prediction windows, with more modalities, and over repeated pre-training runs.

Together, these results suggest that latent prediction combined with bottleneck routing provides a scalable, task-agnostic foundation for multi-modal clinical modeling.

AI USE STATEMENT

We used generative AI tools to format and edit the manuscript, and to generate and debug code for experimental runs. The authors reviewed and verified all AI-assisted text and code, and take full responsibility for the content of this work.

ETHICS STATEMENT

This work uses MIMIC-IV and eICU. Both are de-identified databases, accessed through PhysioNet under credentialed access and the corresponding data use agreements. No new data were collected, and we made no attempt to re-identify individuals. The models are research artifacts and are not validated for clinical use. In particular, the mortality results come from complete records and do not measure early-warning performance.

REPRODUCIBILITY STATEMENT

Cohort definitions and endpoint statistics are given in Appendix A. Label definitions and ICD mappings are given in Appendix B. Architectures, hyperparameters and probe settings are given in Appendices C and D.

REFERENCES

Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

Mido 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, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 1298–1312. PMLR, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

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.

Jiayi Chen and Aidong Zhang. HGMF: Heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1295–1305, 2020.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

L. Gong, Y. Liu, L. Sun, Y. Bi, J. Liu, and X. Zhu. Embracing aleatoric uncertainty in medical multimodal learning with missing modalities. *arXiv preprint arXiv:2601.21950*, 2026.

Rolf HH Groenwold. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*, 4(1):1–7, 2020.

Huan He, William Hao, Yuanzhe Xi, Yong Chen, Bradley Malin, and Joyce C. Ho. A flexible generative model for heterogeneous tabular EHR with missing modality. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-MIWAE: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2006.12871.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9938–9964. PMLR, 2022.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV Clinical Database Demo. *PhysioNet*, January 2023. doi: 10.13026/dp1f-ex47. URL https://doi.org/10.13026/dp1f-ex47. Version 2.2.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022. doi: 10.1038/s41746-022-00712-8.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 1, 2022.

Zihan Liang, Ziwen Pan, and Ruoxuan Xiong. Causal representation learning from multimodal clinical records under non-random modality missingness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28779–28796, Suzhou, China, 2025. Association for Computational Linguistics.

Zachary C Lipton, David Kale, and Randall Wetzel. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pp. 253–270. PMLR, 2016.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 14200–14213, 2021.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689–696, 2011.

Zhiliang Peng, Li Dong, Hangbo Bao, Furu Wei, and Qixiang Ye. A unified view of masked image modeling. *Transactions on Machine Learning Research*, 2023.

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*, 5:180178, 2018. doi: 10.1038/sdata.2018.178.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15638–15650, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024.

Jiaxuan You, Xiaobai Ma, Daisy Yi Ding, Mykel J. Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19075–19087, 2020.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M^{3}care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022.

Xiaoguang Zhu, Lianlong Sun, Yang Liu, Pengyi Jiang, Uma Srivatsa, Nipavan Chiamvimonvat, and Vladimir Filkov. Causal debiasing medical multimodal representation learning with missing modalities. *arXiv preprint arXiv:2509.05615*, 2025.

A COHORTS AND ENDPOINTS

A.1 MIMIC-IV

The MIMIC-IV cohort contains 81,664 adult hospital admissions. It is split by patient into a training set of 65,331 admissions (80%) and a held-out test set of 16,333 admissions (20%). No patient appears in both sets. Each admission has four modalities: ICD-10 procedure codes, medication administrations, radiology reports and demographics (age and sex). Radiology reports are missing in 43.9% of admissions. All other modalities are always present. Input sequences cover the admission up to discharge. Discharge summaries, post-mortem notes and retrospective billing flags are removed. Diagnosis codes are not used as input. They are only used to define the phenotype labels (Appendix B).

Table 4 gives the distribution of the two endpoints. Table 5 compares admissions with and without radiology reports. Tables 6 and 7 give the labels that are only used for the clustering evaluation.

Table 4: MIMIC-IV endpoints: number of admissions (percentage). Phenotyping is multi-label, so the phenotype counts exceed the number of admissions.

Endpoint	**Train** ($N = 65{,}331$)	**Test** ($N = 16{,}333$)	**Total** ($N = 81{,}664$)
In-hospital mortality			
Died	3,434 (5.3)	910 (5.6)	4,344 (5.3)
Survived	61,897 (94.7)	15,423 (94.4)	77,320 (94.7)
Phenotype group			
Endocrine & Metabolic	49,248 (75.4)	12,406 (76.0)	61,654 (75.5)
Circulatory / Cardiovascular	48,201 (73.8)	12,033 (73.7)	60,234 (73.8)
Digestive / Gastrointestinal	35,381 (54.2)	8,925 (54.6)	44,306 (54.3)
Genitourinary / Renal	29,271 (44.8)	7,297 (44.7)	36,568 (44.8)
Mental & Behavioral	27,720 (42.4)	6,965 (42.6)	34,685 (42.5)
Nervous System	26,627 (40.8)	6,763 (41.4)	33,390 (40.9)
Respiratory	26,261 (40.2)	6,558 (40.2)	32,819 (40.2)
Musculoskeletal & Rheumatic	21,594 (33.1)	5,460 (33.4)	27,054 (33.1)
Infectious & Parasitic	17,906 (27.4)	4,473 (27.4)	22,379 (27.4)
Neoplasms (Oncology)	16,721 (25.6)	4,236 (25.9)	20,957 (25.7)
No phenotype label	3,094 (4.7)	762 (4.7)	3,856 (4.7)

Table 5: MIMIC-IV admissions by radiology availability, over the full cohort.

Subset	**Share of admissions**	**Mortality**	**No phenotype label**
All admissions	100%	5.32%	4.07%
All modalities present	56.1%	4.80%	1.02%
Radiology missing	43.9%	5.91%	7.96%

Table 6: Fine-grained diagnostic subgroups, used only for the clustering evaluation.

Diagnostic subgroup	**Train** ($N = 65{,}331$)	**Test** ($N = 16{,}333$)	**Total** ($N = 81{,}664$)
Musculoskeletal & Rheumatic			
Joint & Spine (Mechanical)	15,023	3,761	18,784
Acute Bone & Soft Tissue	8,649	2,240	10,889
Systemic Autoimmune	1,251	342	1,593
Missing / Unassigned	44,111	10,955	55,066
Cardiovascular			
Ischemic & Structural Heart	21,115	5,230	26,345
Hypertensive & Rheumatic	15,169	3,718	18,887
Peripheral Vascular & Venous	8,402	2,169	10,571
Cerebrovascular	3,382	886	4,268
Missing / Unassigned	17,263	4,330	21,593

Table 7: Subject-level hospital readmission in the MIMIC-IV test set, used only for the clustering evaluation.

Readmission	**Test subjects** ($N = 8{,}828$)
Readmitted	4,549 (51.5%)
Not readmitted	4,279 (48.5%)

A.2 eICU

The eICU cohort contains 188,094 ICU stays. We include all stays with at least one modality besides demographics. The cohort is split by patient into a training set of 150,475 stays and a held-out test set of 37,619 stays. Each stay has four modalities: ICD diagnosis codes, medications, entries from the notes table and demographics (age and sex). eICU contains no procedure codes. Input sequences cover the stay from admission to discharge. The in-hospital mortality prevalence is 9.16% in the full cohort and 9.37% in the test set. In the test set, 347 stays have no mortality label.

Table 8 lists the missingness patterns. 65.4% of stays have all modalities. Demographics are always present.

Table 8: eICU missingness patterns. Letters denote the present modalities: a = diagnosis codes, b = medications, c = notes, d = demographics. Mortality is computed over the labelled test stays.

Present	**Missing**	**Cohort stays (%)**	**Test stays**	**Test deaths**	**Test mortality**
abcd	–	123,097 (65.44)	24,734	2,510	10.26%
acd	medications	29,969 (15.93)	6,013	658	11.04%
bcd	codes	21,472 (11.42)	4,192	156	3.73%
bd	codes, notes	8,659 (4.60)	1,716	118	6.92%
cd	codes, medications	3,250 (1.73)	648	23	3.58%
abd	notes	1,130 (0.60)	214	19	8.88%
ad	medications, notes	517 (0.27)	102	8	7.84%
One modality missing		52,571 (27.95)	10,419	833	8.04%
Two modalities missing		12,426 (6.61)	2,466	149	6.08%
All stays		188,094 (100)	37,619	3,492	9.37%

B Phenotype labels and ICD mappings

Phenotyping uses ten diagnostic groups (Table 9). The groups are derived from each admission's ICD-9 and ICD-10 diagnosis codes, mapped to their physiological system. An admission can belong to several groups. Scores are macro-averaged over the ten groups. Admissions without a phenotype label are excluded from the phenotyping metrics. Tables 10 and 11 define finer subgroups within two specialties. Only the cardiovascular subgroups are used in the main text, for the clustering evaluation.

C Model and training details

C.1 Uni-modal encoders

Table 12 summarizes the encoders. The code and medication encoders are BERT models with one code per token and discrete gap tokenization. The text encoder is a BERT-style transformer over TF-IDF token sequences. All three are pre-trained with masked language modeling. The demographic encoder is a deterministic MLP that maps age and sex to two tokens. Each dataset has its own encoders. They are pre-trained on the training set only, before multi-modal training. All encoders are frozen during multi-modal training.

Table 9: Phenotype groups: ten diagnostic categories.

Group	ICD-9	ICD-10	Typical presentations
Infectious & Parasitic	001–139	A00–B99	Sepsis, viral pneumonias, intestinal infections, HIV/AIDS.
Neoplasms (Oncology)	140–239	C00–D49	Malignant tumors, leukemias, benign neoplasms.
Endocrine & Metabolic	240–279	E00–E89	Diabetes mellitus, malnutrition, severe dehydration, thyroid disorders.
Mental & Behavioral	290–319	F00–F99	Delirium, dementia, alcohol withdrawal syndrome, schizophrenia.
Nervous System	320–389	G00–G99	Epilepsy, Parkinson's disease, Alzheimer's disease, meningitis.
Circulatory / Cardiovascular	390–459	I00–I99	Heart failure, acute myocardial infarction, arrhythmias, hypertension.
Respiratory	460–519	J00–J99	COPD exacerbations, asthma, bacterial pneumonia, pleural effusion.
Digestive / Gastrointestinal	520–579	K00–K93	Gastrointestinal hemorrhage, liver cirrhosis, acute pancreatitis, appendicitis.
Musculoskeletal & Rheumatic	710–739	M00–M99	Rheumatoid arthritis, osteoarthritis, systemic lupus, osteomyelitis.
Genitourinary / Renal	580–629	N00–N99	Acute kidney injury, chronic kidney disease, urinary tract infections.

Table 10: Musculoskeletal & Rheumatic subgroups.

Subgroup	ICD-9	ICD-10	Typical presentations
Systemic Autoimmune	710	M30–M36	Multi-organ involvement (lupus, scleroderma) and chronic immunosuppression.
Joint & Spine (Mechanical)	711–724	M00–M25, M40–M54	Arthropathies and dorsopathies (osteoarthritis, gout, rheumatoid arthritis, spinal stenosis).
Acute Bone & Soft Tissue	725–739	M60–M94	Osteomyelitis, rhabdomyolysis and pathological fractures.

C.2 Optimization

All multi-modal models are trained with AdamW and a cosine learning-rate schedule with linear warm-up (Table 13). The frozen encoders are excluded from the optimizer.

C.3 MLPA

Table 14 lists the MLPA settings. MLPA has 1.9M trainable parameters on MIMIC-IV.

C.4 MAE

The MAE uses the same encoders, projections, empty-state tokens, masking and MBT as MLPA. It replaces the latent predictors with transformer decoders that reconstruct the masked inputs. The reconstruction losses are cross-entropy for discrete codes, mean squared error for TF-IDF vectors, and a combination of mean squared error and binary cross-entropy for demographics. Table 15 lists the settings that differ from MLPA. The MAE has 3.3M trainable parameters on MIMIC-IV.

C.5 MUSE+

MUSE+ (Wu et al., 2024) represents each record as a patient node connected to the nodes of its observed modalities. A missing modality has no edge. During training, MUSE+ randomly drops edges to create augmented views for its contrastive objective. It is trained with the endpoint labels, and uses records without a label as unlabeled patients. A separate model is trained for each

Table 11: Cardiovascular subgroups.

Subgroup	ICD-9	ICD-10	Typical presentations
Hypertensive & Rheumatic	390–405	I00–I15	Rheumatic fever, chronic rheumatic heart disease, essential and secondary hypertension.
Ischemic & Structural Heart	410–429	I20–I52	Acute myocardial infarction, heart failure, arrhythmias, valvular disease.
Cerebrovascular	430–438	I60–I69	Ischemic stroke, intracerebral hemorrhage, transient ischemic attack.
Peripheral Vascular & Venous	440–459	I70–I99	Deep vein thrombosis, aortic aneurysm, peripheral artery disease.

Table 12: Uni-modal encoders. The BERT encoders are pre-trained with masked language modeling. All encoders are frozen during multi-modal training. Sources are given as MIMIC-IV / eICU.

Modality	Source	Tokenization	Encoder	Max. len.
Codes	Procedures / diagnoses	Gap tokens	BERT ($d = 128$)	120
Medications	Medication codes	Gap tokens	BERT ($d = 128$)	512
Text	Radiology / notes	TF-IDF	BERT ($d = 128$)	40
Demographics	Age and sex	Direct mapping	MLP ($d = 64$)	2

endpoint. All MIMIC-IV admissions have a mortality label, so no unlabeled patients are used for this endpoint. For MIMIC-IV phenotyping and eICU mortality, records without a label are used as unlabeled patients. MUSE+ has 0.8M trainable parameters on MIMIC-IV. Table 16 lists its settings.

In the main experiments, MUSE+ receives the same frozen uni-modal embeddings as MLPA. We also trained a variant on MIMIC-IV in which the uni-modal encoders are fine-tuned end-to-end, as in the original work. This variant has about 0.4M parameters. Table 17 compares the two variants. Mortality results and phenotyping AUPRC are similar. Fine-tuning raises the phenotyping AUROC (0.835 vs. 0.817).

C.6 Perceiver routing variant

The Perceiver variant replaces the MBT with latent cross-attention (Jaegle et al., 2022). A set of learned latent queries attends over the concatenated tokens of all streams. The latents are mean-pooled into Z_f. All other components and settings are the same as for MLPA. The latent query dimension 128 with 2 layers to match the MBT fusion layers.

D Evaluation protocol

Probes. MLPA and MAE representations are evaluated with a linear probe and a two-layer MLP probe (Table 18). We use 10-fold cross-validation on the held-out test set. In each fold, the probe is trained on nine folds for 100 epochs, without early stopping, and evaluated on the remaining fold. Records without a label for an endpoint are excluded from that endpoint's metrics. We report the mean and standard deviation over the ten folds. MUSE+ is trained on the full training set and evaluated on the same test folds. We compare models with a paired Wilcoxon signed-rank test over the ten folds.

Subgroup evaluation. On MIMIC-IV, the probes for MAE and MLPA are fit and cross-validated within each subgroup. Both subgroups are large, with about 8,200 and 6,500 training admissions per fold. On eICU, several missingness patterns are small. The probes are therefore fit once on all test stays, and their out-of-fold predictions are scored within each subgroup. For the merged eICU groups in Table 1, the metrics are computed on the pooled out-of-fold predictions. In both datasets, MUSE+ is trained once on the full training set and evaluated on each subgroup.

Table 13: Optimization settings shared by all multi-modal models.

Setting	Value
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Weight decay	0.01
Learning-rate schedule	Cosine with linear warm-up
Warm-up	5% of total steps
Peak learning rate	1×10^{-4}
Gradient clipping	$\|g\|_2 \leq 1.0$
Batch size (train / validation)	64 / 32

Table 14: MLPA settings.

Setting	Value
Shared dimension d	128
MBT layers L	6
Modality-specific layers $L - n$	4
Fusion layers n	2
MBT feed-forward dimension	128
Bottleneck tokens K	8
MBT attention heads	4
Predictor layers	2
Predictor attention heads	2
Target masking ratio	60%
Loss weights (codes, medications, text, demographics)	1.0, 1.0, 1.0, 1.0
Epochs	200

E eICU RESULTS BY MISSINGNESS PATTERN

Tables 19 and 20 report mortality results for each eICU missingness pattern. The metrics are computed per fold and averaged. The three smallest patterns have 23, 19 and 8 deaths in the test set. Their fold standard deviations are large, and we do not interpret them individually.

F CLUSTERING EVALUATION

Davies–Bouldin index. Let the embeddings be partitioned into groups $G_1, \ldots, G_J$ by a label. Let μ_j be the centroid of group j, and s_j the mean Euclidean distance of its members to μ_j. The Davies–Bouldin index (Davies & Bouldin, 1979) is

$$\mathrm{DBI} = \frac{1}{J} \sum_{j=1}^{J} \max_{j' \neq j} \frac{s_j + s_{j'}}{\|\mu_j - \mu_{j'}\|_2}. \tag{9}$$

Lower values mean more compact and better separated groups.

Labels. Mortality uses the two outcome classes. For the phenotype groups, an admission that belongs to several groups is assigned the group of its highest-priority diagnosis, as ranked in MIMIC-IV. Readmission is defined per patient on the test set (Table 7). A patient is readmitted if a later admission starts after the end of their first admission. For readmission, the earliest embedding represents the patient. The cardiovascular subgroups follow Table 11. Admissions without a label, and admissions without a cardiovascular diagnosis for the cardiovascular subgroups, are excluded.

Embeddings. For MLPA and MAE, we use Z_f on the MIMIC-IV test set. For MUSE+, we use two models, one trained for mortality and one trained for phenotyping. Their patient-node embeddings are extracted in evaluation mode, without edge dropout, and L2-normalized (Algorithm 1).

Table 15: MAE settings that differ from MLPA.

Setting	Value
Decoder layers	4
Decoder attention heads	4
Target masking ratio	60%
Loss weights (codes, medications, text, demographics)	1.0, 1.0, 1.0, 1.0
Epochs	200

Table 16: MUSE+ settings.

Setting	Value
GNN hidden dimension	128
GNN layers	2
Dropout	0.15
Contrastive temperature τ	0.05
Maximum epochs	200

G Cohort Selection and Implementation Details

G.1 Cohort Selection and Inclusion Criteria

MIMIC-IV: We extracted hospital admissions from the MIMIC-IV database (v2.2) (Johnson et al., 2023). We restricted the cohort to adult patients (age $\geq$ 18 at admission) with recorded lengths of stay of at least 24 hours. Records lacking valid admission or discharge timestamps, or those missing administrative demographics, were excluded. To prevent target leakage, all retrospective billing flags, discharge summaries, and post-mortem notes were stripped from the record prior to sequence generation.

eICU Collaborative Research Database: We selected ICU stays from the multi-center eICU platform (v2.0) (Pollard et al., 2018) corresponding to adult admissions ($\geq$ 18 years). For patients with multiple ICU stays during a single hospitalization, only the initial stay was retained to prevent patient-state autocorrelation. Stays with unresolvable hospital discharge statuses or admission offsets outside valid recording windows were filtered out.

G.2 Sequence Framing and Target Leakage Controls

Input sequences encompass clinical events from admission up to discharge. For non-surviving patients, the observation window terminates strictly at the recorded timestamp of death. Consequently, in-hospital mortality is evaluated as a retrospective stay classification task comparing complete trajectories under identical sequence boundaries, rather than a prospective early-warning forecast. Diagnostic ICD codes used to construct the multi-label phenotyping targets on MIMIC-IV are derived exclusively from final discharge billing records and are excluded from the input sequence.

G.3 Modality Specifications and Missingness Dynamics

The four input streams are structured as follows:

- **Clinical Codes:** In MIMIC-IV, this stream contains ICD-10-PCS procedure codes logged during the stay. In eICU, procedural codes are absent; thus, the code stream consists of ICD-9 and ICD-10 clinical diagnosis codes.
- **Medications:** Coded medication administrations and prescriptions extracted from the respective pharmacy and eMAR tables.
- **Clinical Text:** In MIMIC-IV, free-text reports are drawn exclusively from diagnostic radiology studies. In eICU, text entries are extracted from the routine care provider notes table.
- **Demographics:** Baseline patient age at admission and biological sex.

Table 17: MUSE+ with frozen and with fine-tuned uni-modal encoders on MIMIC-IV. Mean$_{\pm \text{s.d.}}$ over 10 folds. The fine-tuned variant was not trained on eICU.

	Phenotyping		In-hospital mortality	
Encoders	AUPRC	AUROC	AUPRC	AUROC
All test admissions				
Frozen	$0.823_{\pm.003}$	$0.817_{\pm.003}$	$0.672_{\pm.042}$	$0.952_{\pm.012}$
Fine-tuned	$0.821_{\pm.003}$	$0.835_{\pm.003}$	$0.662_{\pm.039}$	$0.956_{\pm.010}$
All modalities present				
Frozen	$0.831_{\pm.005}$	$0.815_{\pm.005}$	$0.638_{\pm.095}$	$0.938_{\pm.018}$
Fine-tuned	$0.828_{\pm.005}$	$0.832_{\pm.005}$	$0.618_{\pm.090}$	$0.940_{\pm.019}$
Radiology missing				
Frozen	$0.816_{\pm.005}$	$0.819_{\pm.004}$	$0.706_{\pm.065}$	$0.962_{\pm.014}$
Fine-tuned	$0.815_{\pm.004}$	$0.837_{\pm.004}$	$0.712_{\pm.061}$	$0.967_{\pm.009}$

Table 18: Probe settings.

Setting	**Value**
Probes	Linear layer; two-layer MLP
MLP hidden units	128
MLP dropout	0.2
Optimizer	AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4})
Batch size	1024
Epochs	100 (no early stopping)

Missingness patterns differ substantially across institutions:

- **MIMIC-IV:** Missingness is isolated to radiology text, where 43.9% of admissions lack imaging studies (representing patients who did not undergo radiological evaluation). Demographics, procedure codes, and medications are present across all admissions in the cohort.
- **eICU:** Modality absence occurs across six discrete permutations, affecting 34.6% of all stays. Specifically, diagnosis codes are absent in 17.7% of stays, medication records are absent in 17.9%, and clinical notes are missing in 5.5%. Demographics are complete across all stays.

During inference on incomplete stays, missing modalities are substituted with learned, modality-specific empty-state embeddings.

G.4 Uni-Modal Encoder Architecture and Pre-Training

To ensure fair comparison across multi-modal fusion heads, all uni-modal encoders are pre-trained independently and frozen:

- **Structured Streams (Codes and Medications):** Formatted as discrete sequences tokenized with one code per token. Encoders follow a 4-layer BERT architecture pre-trained via masked language modeling (MLM) with a 15% masking rate exclusively on the respective training partitions.
- **Clinical Text:** Tokenized and encoded using domain-specific BERT models initialized from clinical checkpoints and adapted via MLM on the training split notes.
- **Demographics:** Age is normalized to zero mean and unit variance; sex is binary-encoded. Both are projected into the common model embedding dimension via a deterministic linear transformation.

All unimodal encoders remain completely frozen during downstream fusion pre-training and probing evaluations.

Table 19: eICU mortality AUPRC by missingness pattern. Mean$_{\pm \text{s.d.}}$ over 10 folds. Letters denote the present modalities (Table 8).

Present	Missing	Stays	Deaths	MAE	MUSE+	MLPA
abcd	–	24,734	2,510	$0.415_{\pm.037}$	$0.408_{\pm.043}$	$0.458_{\pm.039}$
acd	medications	6,013	658	$0.384_{\pm.047}$	$0.384_{\pm.029}$	$0.395_{\pm.046}$
bcd	codes	4,192	156	$0.253_{\pm.062}$	$0.263_{\pm.104}$	$0.281_{\pm.076}$
bd	codes, notes	1,716	118	$0.222_{\pm.052}$	$0.393_{\pm.121}$	$0.250_{\pm.066}$
cd	codes, medications	648	23	$0.340_{\pm.338}$	$0.403_{\pm.302}$	$0.247_{\pm.230}$
abd	notes	214	19	$0.295_{\pm.214}$	$0.610_{\pm.326}$	$0.323_{\pm.222}$
ad	medications, notes	102	8	$0.318_{\pm.314}$	$0.546_{\pm.430}$	$0.338_{\pm.337}$

Table 20: eICU mortality AUROC by missingness pattern. Mean$_{\pm \text{s.d.}}$ over 10 folds.

Present	Missing	Stays	Deaths	MAE	MUSE+	MLPA
abcd	–	24,734	2,510	$0.843_{\pm.012}$	$0.820_{\pm.015}$	$0.859_{\pm.012}$
acd	medications	6,013	658	$0.817_{\pm.028}$	$0.798_{\pm.023}$	$0.826_{\pm.020}$
bcd	codes	4,192	156	$0.797_{\pm.046}$	$0.784_{\pm.061}$	$0.818_{\pm.043}$
bd	codes, notes	1,716	118	$0.769_{\pm.076}$	$0.779_{\pm.070}$	$0.792_{\pm.075}$
cd	codes, medications	648	23	$0.760_{\pm.188}$	$0.844_{\pm.166}$	$0.675_{\pm.283}$
abd	notes	214	19	$0.658_{\pm.141}$	$0.862_{\pm.108}$	$0.793_{\pm.074}$
ad	medications, notes	102	8	$0.766_{\pm.224}$	$0.744_{\pm.267}$	$0.789_{\pm.168}$

H NOTATION

Table 21 summarizes the notation of Section 3. The operator $\|$ always denotes concatenation along the sequence dimension.

Table 21: Notation.

Symbol	Sec.	Meaning
M, m	3.1	Number of modalities; modality index
X_m	3.1	Raw input of modality m
E_m	3.1	Frozen encoder of modality m
P_m	3.1	Learned linear projection to dimension d
U_m	3.1	Projected sequence, $U_m \in \mathbb{R}^{T_m \times d}$
T_m	3.1	Sequence length of modality m
d	3.1	Shared dimension
CLS_m	3.1	Learned empty-state token of modality m
Z_m	3.1	Stream of modality m, $Z_m = [\text{CLS}_m \,\|\| \, U_m]$
L, n	3.2	Number of MBT layers; number of fusion layers
K	3.2	Number of bottleneck tokens
$B^{(l)}$	3.2	Shared bottleneck tokens after layer l, $B^{(l)} \in \mathbb{R}^{K \times d}$
$\hat{B}_m^{(l)}$	3.2	Bottleneck candidate of stream m at layer l
$b_k^{(L)}$	3.2	k-th terminal bottleneck token
Z_f	3.2	Fused representation, $Z_f \in \mathbb{R}^d$
$\mathcal{C}_m$, $\mathcal{T}_m$	3.3	Context and target positions of modality m
Z_m^t, $z_{m,i}^t$	3.3	Target sequence $E_m(X_m)$; target at position i
Z_m^{ctx}	3.3	Context stream
$\mathcal{O}$	3.3	Set of observed modalities
Pred_m	3.3	Predictor of modality m
$E_{\text{mask}}^{(m)}$	3.3	Positional mask tokens, $E_{\text{mask}}^{(m)} \in \mathbb{R}^{\lvert\mathcal{T}_m\rvert \times d}$
$H_{\text{pred},m}$	3.3	Predictor input, $[Z_f \,\|\| \, E_{\text{mask}}^{(m)}]$
$\hat{z}_{m,i}$	3.3	Prediction at target position i
$\mathcal{L}$	3.3	Training loss

Algorithm 1 MUSE+ patient embedding extraction

Require: Uni-modal embeddings $\mathbf{E} = \{E_a, E_b, E_c, E_d\}$
Require: Missingness indicators $\mathbf{M} = \{M_a, M_b, M_c, M_d\}$
Require: Trained graph neural network GNN_θ

GNN_θ.eval() ▷ Disable dropout and other training-only modules
drop_edges $\leftarrow$ False ▷ No edge dropout
$\mathbf{Z} \leftarrow \text{GNN}_\theta(\mathbf{E}, \mathbf{M}, \text{drop_edges})$
$\mathbf{Z}_{\text{out}} \leftarrow \mathbf{Z}/\|\mathbf{Z}\|_2$ ▷ L2 normalization along the feature dimension
return $\mathbf{Z}_{\text{out}}$